(2)

IMACS '91

13TH WORLD CONGRESS ON COMPUTATION AND APPLIED MATHEMATICS

JULY 22 - 26, 1991
TRINITY COLLEGE DUBLIN
IRELAND

PROCEEDINGS

IN FOUR VOLUMES

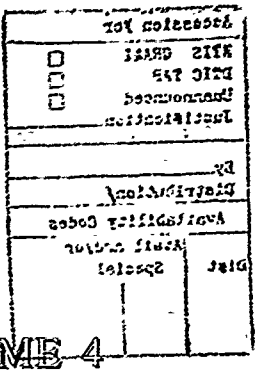
VOLUME 4

IMACS '91

Proceedings of the 13th IMACS World Congress on Computation and Applied Mathematics

July 22-26, 1991, Trinity College, Dublin, Ireland

in four volumes

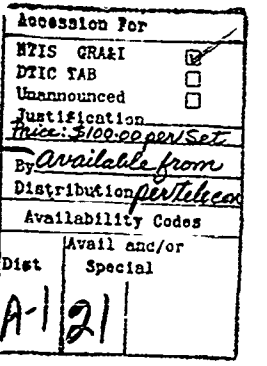


VOLUME 4

Modelling and Simulation for Electrical, Electronic and Semiconductor Devices
Computation for Management Systems
Applications of Modelling and Simulation
Environmental Systems Simulation
Software Forum
Poster Sessions
Author Index

EDITED BY: R Vichnevetsky
Rutgers University
New Brunswick, USA

J J H Miller
Trinity College
Dublin, Ireland



IMACS Symposium Rutgers Univ Dept of Computer Science New Brunswicks, NJ 08903

100.00 per set 4 Vols.

Printed in Ireland by Criterion Press, Dublin

MODELLING AND SIMULATION OF INVERTER FED SLOTLESS PERMANENT MAGNET MACHINES

B.NOGAREDE, A.D. KONE, M.LAJOIE MAZENC

Laboratoire d'Electrotechnique et d'Electronique Industrielle, Unité de Recherche Associée au CNRS n°847, ENSEEIHT, 31071 Toulouse, France.

Abstract: The authors present a modelling method allowing the study of voltage inverter fed slotless permanent magnet machines. The suggested method is based on an analytical field calculation inside the magnetic structure which is associated with a numerical solution of the global electric equation.

INTRODUCTION:

For special applications needing the lowest vibration level possible, the slotless permanent magnet machine fed by a pulse width modulation controlled voltage inverter, seems to be a most efficient solution [1]. In this article, a modelling and simulation method adapted to such a drive system is presented. In order to take into account the power subdivision required by high powered drive, the general case of a multistar and polyphased system is considered.

I - GENERAL ELECTRIC EQUATION:

The general electric scheme of the studied structure, which is made up of a e-star m-phased armature, is given on Figure 1.

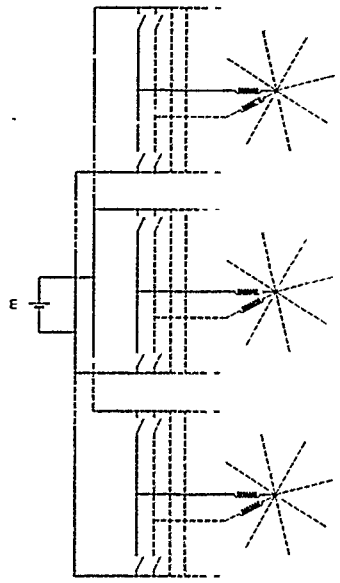


Figure 1 - General electric scheme.

The different stars, electrically independent and magnetically coupled are fed by e inverters including m phases. The switches of a same inverter leg are unidirectional for voltage and bidirectional for current.

Neglecting the voltage switching duration, the applied voltages on the machine different phases are entirely defined by the inverter working sequences. In the case of the mecanic steady state (constant rotating speed) the system behaviour is entirely described by the electric equation which is given in a matrical form for all the machine phases by:

$$[V] = [R][I] + \frac{d}{dt}[\Phi] \qquad (1)$$

The unicolumn vectors [V], [I] and [Φ], including e.m components, respectively correspond to voltage, current and flux in the stator armature. The diagonal matrix [R], corresponds to the winding resistance.

The determination of the currents from Equation (1) knowing the applied voltages, requires the calculation of flux [Φ]. An expression of [Φ] can be obtained from an analytical field calculation in the machine magnetic structure.

II - MAGNETIC STRUCTURE MODELLING:

In the case of a smooth pole and slotless stator machine, the magnetic structure can be modelled by a two-dimensional analytical field calculation method in a (r,θ) co-ordinates system. As shown on Figure 2, the studied structure can be divided into several concentric layers corresponding to the magnetic materials and the field sources (magnets or currents).

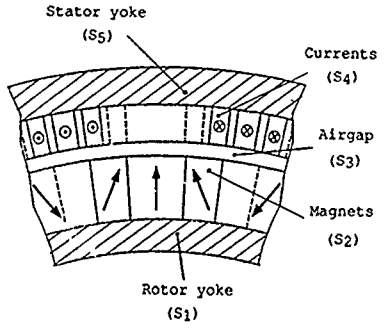


Figure 2 - Magnetic structure.

To the extent that the thickness and the permeability of these layers are constant, an analytical solution of Poisson's equation can be obtained in the different (S_k) zones. The development into Fourier's series of the fonctions defining the magnet and current spatial distribution allows the determination of the vector potential expression in every zone [2].

The vector potential non nil compcnent, for a k suffix layer, is given by:

$$a_k = \sum_{g=-\infty}^{+\infty} [X_{gk} r^{|g|p} + Y_{gk} r^{-|g|p} + Z_{gk}(r)] e^{jgp\theta} \qquad (2)$$

where the coefficient X_{gk}, Y_{gk} and the Z_{gk} functions are calculated from the structure's geometrical and electromagnetic characteristics, p is the pole pair number.

Elaborated magnet arrangements and winding structures can be taken into account in a quite simple way by this method: multiblock poles with a magnetization either radial or parallel to any direction, multistar polyphased windings with complex distributions of conductors...

III - FLUX DETERMINATION:

The different electromagnetic quantities defining the working of the machine can be analytically determined by the above vector potential expression. Particularly the flux in any winding phase (nth phase of the hth star) is given by:

$$\Phi_{hn}(t) = L. \int_{(S_4)} A_4(r,\theta,t).C_{hn}(\theta)\, r d\theta dr \qquad (3)$$

where A_4, function of the space variables r, θ and of the time t, represents the vector potential norm in the currents' layer. $C_{hn}(\theta)$ function represents the angular distribution of the considered phase conductors, L is the axial length of the machine.

By substituting the Fourier's series defining A_4 and C_{hn} in Equation (2), and after integration, the flux can be put in the form:

$$\Phi_{hn}(t) = \Phi_{A_{hn}}(t) + \sum_{i=1}^{e} \sum_{l=1}^{m} M_{hnil}.I_{il}(t) \qquad (4)$$

This flux is divided into a no load flux $\Phi_{A_{hn}}(t)$, corresponding to the magnet contribution, and an armature reaction flux, M_{hnil} representing the mutual inductance between the considered phase and the lth phase of the ith star, in which the $I_{il}(t)$ current flows. The $\Phi_{A_{hn}}(t)$ and M_{hnil} quantities are developed into Fourier's series of which coefficients depend on the geometric and magnetic structure characteristics.

Considering all the machine phases, the flux vector ultimately becomes:

$$[\Phi] = [\Phi_A] + [M][I] + [L_S][I] \qquad (5)$$

where the no load flux $[\Phi_A]$ and the inductance matrix [M] are analytically defined from the magnetic structure modelling. The $[L_S]$ matrix makes it possible to take into account the magnetic leakages in the winding overhangs, which are not considered by the two dimensions modelization. Consequently, the general electric equation (1) which governs the evolution of the currents in the machine windings is perfectly defined and becomes:

$$[V] - \frac{d}{dt}[\Phi_A] = [R][I] + [L_S + M]\frac{d}{dt}[I] \qquad (6)$$

Knowing the voltages applied by the inverter, and the electromotive force $\frac{d}{dt}[\Phi_A]$ determined by its analytical expression, Equation (5) solution makes it possible to determine the currents in the machine and to deduct the space-time evolution of the different electromagnetic quantities (flux density, torque, force density in the winding...) from the modelization of the magnetic structure.

IV - ELECTRIC EQUATION SOLUTION METHOD:

Equation (5) can be put in state form:

$$\frac{d}{dt}[X] = [A][X] + [B][U] \qquad (7)$$

with:

$$[X]=[I],\ [U]=[V]-\frac{d}{dt}[\Phi_A],\ [B]=[L_S+M]^{-1},$$
$$[A]=-[B].[R]$$

To the extent that [A] and [B] matrix are independent of time and of vectors [X] and [U], a suitable method for the solution of Equation (6) is the matrix exponential method [3]. The general solution of Equation (6) then takes the form of:

$$[X(t)] = e^{[A](t-t_0)}.[X(t_0)] + \int_{t_0}^{t} e^{[A](t-\tau)}.[B].[U(\tau)]\, d\tau \qquad (8)$$

which, if the calculation step is chosen in such a way that the vector [U] remains constant during this period, leads to the exact recurrent analytical expression of the solution:

$$[X(t_0+h)] = e^{[A]h}[X(t_0)] + [A]^{-1}.(e^{[A]h} - [1]).[B].[U(t_0)] \qquad (9)$$

where [1] is the unitary matrix.

According to the inverter control type adopted, the calculation step can be chosen either variable or constant. So, if the inverter command is a priori known, it is interesting to place the calculation points at the supply voltage switching instants. If these instants are near enough, the e.m.f variations are negligible on that time scale and the command vector [U] can be considered as constant between two switching instants. This solution allows the minimization of the number of the calculation points, but requires an exponential matrix calculation for every new step.

If the switching instants are a priori unknown, which happens in particular when the inverter command comes from a regulation of the currents in the machine, the calculation step needs to be chosen sufficiently fine to limit the switching errors. The time calculation rise due to the increase of the calculation points is limited by the adoption of a constant calculation step which requires only one matrix exponential calculation.

V - EXAMPLE:

To illustrate the proposed method, the case of a high powered four-star three-phased sinusoïdal waveform structure has been considered. The four PWM three-phased inverters being controlled by a sinusoïdal modulation, the current waveforms obtained in two different configurations of the structure has been studied.

The first configuration corresponds to a in-phase four-star winding and to a synchronized control of the four inverters. As shown on Figure 3, the phase current ripple is rather weak, in as much as the frequency modulation here is 100 times higher than the fundamental frequency.

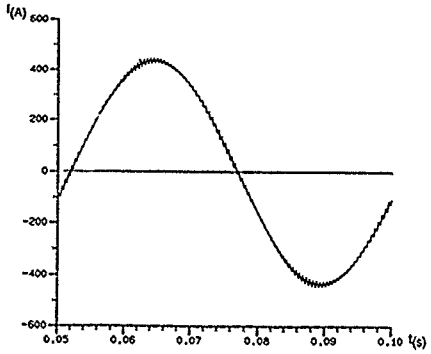


Figure 3 - Phase current, synchronized control.

The second configuration corresponds to the case where the four stars are spatialy shifted from one another. The fundamental currents of the different stars having to be shifted too, the control of the four inverters here is such that the voltage switchings in the different stars are not simultaneous. So, during a star voltage switching, the other stars, connected to the direct voltage source, act as an amortisseur winding. The transient inductance which corresponds to a leakage inductance between the different stars is noticeably weaker than the synchronous inductance involved in the previous configuration. As shown on Figure 4, the current ripple obtained is far higher.

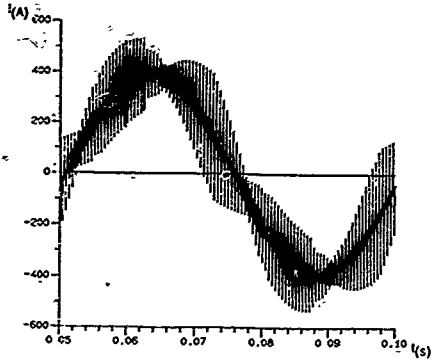


Figure 4: Phase current, shifted control.

CONCLUSION:

In this paper, a modelling and simulation method based on an analytical field calculation associated with a matrix exponential solution of the electric equation has been presented. The proposed method, which considers the general case of a multistar polyphased system, allows the taking into account of quite complex inductor and winding structures. By adopting a constant or variable calculation step according to the inverter working type, the matrix exponential solution technique enables the minimization of the calculation times. Moreover, thanks to its analytical nature, by showing the influence of the different design parameters, the proposed modelization method has turned out to be a particularly performing optimization tool.

REFERENCES:

[1] B.NOGAREDE, "Etude de moteurs sans encoches à aimants permanents de forte puissance à basse vitesse", thèse de Doctorat de l'Institut National Polytechnique, Toulouse, 1990.

[2] B.NOGAREDE, M LAJOIE MAZENC, B.DAVAT, "Modélisation analytique des machines à aimants à induit sans encoches" Revue de Physique Appliquée, 25 (1990), 707-720.

[3] J.SCHONFK, "Simulation numérique de convertisseurs statiques", thèse de Docteur-ingénieur de l'Université Paul Sabatier, Toulouse, 1977.

MODELLING AND SIMULATION OF INDUCTION GENERATOR WITH CURRENT SOURCE INVERTER EXCITER

Santana J.
INIC/IST Secção Maq. Eléctricas e Elect. de Potência
Av. Rovisco Pais 1096 Lisboa Codex
PORTUGAL

Margato E.
ISEL/Energia e Sistemas de Potência
Av. Conselheiro Emídio Navarro 1900 Lisboa
PORTUGAL

ABSTRACT The induction generator excited by a current source inverter is analysed. This system can be seen like a DC machine,so the sliding contact circuit theory is used to develop a new dynamical model.

Using that model, the steady-state and transient behaviours are presented. Experimental results are in agree with the theoretical ones.

I. INTRODUCTION

The induction generator excited by a current source inverter is a system proposed by Sato and al in reference [1].

There are two ways to study that system : 1) the machine with its equations is constrained to the conditions given by the electronic circuits, so the overall model results of the compatibility of the relations which govern the components.This was done in reference [1], where the steady-state study was presented , 2) the system is considered a new electrical machine for which it is possible to apply the equation of sliding contact circuits [2].

In this paper, a new dynamical model is obtained with the second method. The steady-state as well as some transient behaviours are presented. The theoretical results are in agree with the experimental ones. So the model is suitable for analysis and synthesis.

II. SYSTEM MODELLING

The experimental system is presented in FIG.1. The induction generator is excited by a current source inverter. The load is a diode bridge supplying a DC circuit.

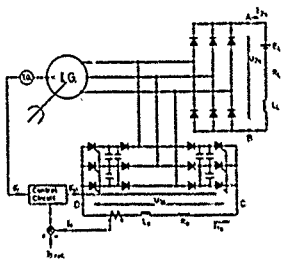


FIG.1 : System representation

The system presented in FIG.1 can be seen like a commutation machine. The current source inverter and the diode bridge are two DC inputs, sliding on a commutator,FIG.2 The rotor windings are represented by α, β circuits.

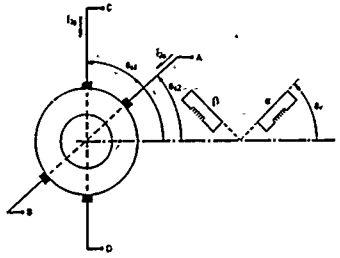


FIG.2: Equivalent system.

Using the sliding contact circuit theory, the stator equations are given by:

$$U_{1s} = r_{1s} i_{1s} + \frac{d\psi_{1s}}{dt} + \frac{0 - \dot{\theta}_{s1}}{i_{1s}} \cdot \frac{\delta W_{cm}}{\delta \theta_{s1}} \qquad (1)$$

$$U_{2s} = r_{2s} i_{2s} + \frac{d\psi_{2s}}{dt} + \frac{0 - \dot{\theta}_{s2}}{i_{2s}} \cdot \frac{\delta W_{cm}}{\delta \theta_{s2}} \qquad (2)$$

Where: W_{cm} is the system co-energy

Ψ_{1s} and Ψ_{2s} are the stator flüxes.

The fictitious brushes of the commutation circuits are not fixed. Their variation laws depend of convert physic.

The position of CD brushes (current inverter) on the commutator is fixed by the control circuit [3]:

$$\dot{\theta}_{s1} = f\left(\dot{\theta}_r, i_{1s}, I_{1ref}\right) \qquad (3)$$

The AB brushes are related with the diode bridge, so they are placed on a commutator position of maximum voltage. To obtain this condition we use a test circuit with voltage U'_{2s} and position θ'_{s2} [4]:

$$\left(\frac{\delta U'_{2s}}{\delta \theta'_{s2}}\right)_{\theta'_{s2} = \theta_{s2}} = 0 \qquad (4)$$

The rotor behaviour is represented by two quadrature windings α,β:

$$0 = r_\alpha i_\alpha + \frac{d\Psi_\alpha}{dt} \tag{5}$$

$$0 = r_\beta i_\beta + \frac{d\Psi_\beta}{dt} \tag{6}$$

Where: Ψ_α and Ψ_β are the rotor fluxes.

The inverter control is done as FIG.3 shows, so in this case equation (3) ,i e, the inverter frequency is given by:

$$\dot{\theta}_{s1} = \dot{\theta}_r + K\left(I_{1ref} - \frac{1}{1+\tau_1 p} i_{1s}\right) \tag{7}$$

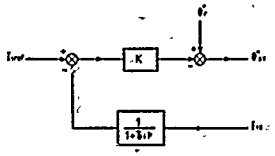


FIG.3-Control circuitt

In normal operation voltage U_{1s} is closely zero and the position of CD brushes is near 90º, the stator windings are not in quadrature to permit the inverter DC current control (excitation).

III. SIMULATION AND EXPERIMENTAL RESULTS

In the following figures, we present some results for the steady-state and the transient behaviour. The theoretical results were obtained with the developed model. These and the experimental data are concerned with standard 2 kW induction machine squirrel- cage.

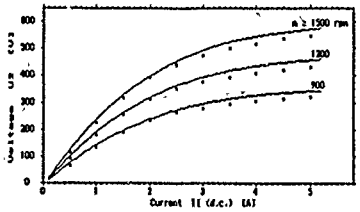


FIG.4 - Unload characteristic.

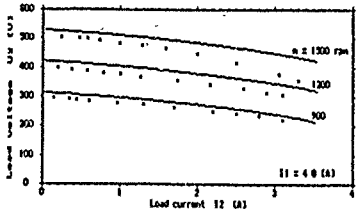


Fig.5 - Output characteristic.

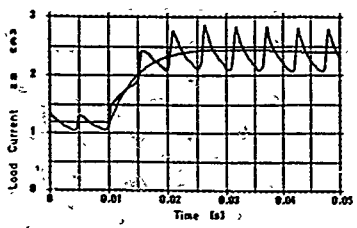


Fig.6 - Output current evolution after a load perturbation (ΔR_1)

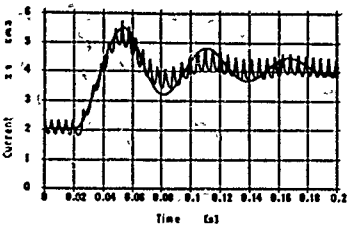


FIG.7 Inverter current evolution after a current reference perturbation (ΔI_{1ref}).

IV. CONCLUSIONS

The equation of sliding contact circuits has been used to study the commutation machines. Also in this paper, this theory was used to obtained a dynamical model of the induction generator with current source inverter exciter.

To validate that model, an experimental system was built. The theoretical results are in agree with the experimental ones. So the developed model is suitable for the analysis and synthesis of experimental systems.

V. REFERENCES

[1] N. Sato, Y. Hayashi, H. Umida
" Load characteristics of induction generator with static exciter "
Proc. ICEM- Athens 1980

[2] M.S. Garrido
" Les équations générales des machines electriques deduites de l'electromagnetisme. "
Revue E, vol. VI, n.4, 1971

[3] J. Santana
" Modelling and stability of the induction motor driven by current source inverter "
Proc. ICEM- Athens 1980

[4] H. Buyse and M.S. Garrido
" A new dynamical model of rectified output AC generator "
Electric Machines and Electromechanics Nº 3, 1978

EQUIVALENT MODELS FOR TUBULAR LINEAR INDUCTION ACTUATORS

CARLOS CABRITA
IST/INIC, DEEC, Secção de Máquinas Eléctricas e Electrónica de Potência
Av. Rovisco Pais - 1096 Lisboa Codex, Portugal.

ABSTRACT - For the analysis and design of the tubular actuator (TA), an equivalent model derived from the similar model used for the low speed flat linear actuator, is presented. To test the theory, two experimental TAs were designed and built by the author. The results are presented and discussed. Using the maximum value of force per power input at standstill, as an optimum design criterion, the optimized design formulas are proposed.

LIST OF SYMBOLS

- D - mean diameter of the rotor aluminium sheet, m
- D_e - maximum diameter of co-axial coils, m
- D_i - minimum diameter of co-axial coils, m
- d_c - standard wire copper diameter, m
- g - actual airgap length, m
- I_1 - r.m.s. phase current, A
- N_1 - number of turns per phase
- p - number of pole pairs
- R_1 - stator winding resistance, Ω
- R_2 - rotor aluminium resistance referred to the stator, Ω
- s - slot width, m
- t - rotor aluminium sheet thickness, m
- X_m - magnetizing reactance, Ω
- z_c - number of turns per coil
- ρ - volume resistivity of the aluminium, Ω.m
- ρ_c - volume resistivity of cooper, Ω.m
- τ_p - pole pitch, m
- τ_s - slot pitch, m
- ω - angular supply frequency, 1/s

1. INTRODUCTION

The small TA is an electric piston, and compared with compressed air and hydraulic systems, is more advantageous because no compressors nor conducts are necessary. Besides, the TA is very reliable and without maintenance. The actual applications of TA are [1, 2, 3]: Pumps, valves, bottle transfers, oscillators and lifts. Other applications are proposed [4]: Sliding doors and curtain operators, robotics, weaver's looms, switch - pointer drives, nuclear reactors and aeronautics.

2. EQUIVALENT MODEL

The TA steady - state operational characteristics may be derived and analysed from a simplified equivalent model per phase shown in fig. 1. The core losses and leakage flux have been neglected. The topology of this equivalent circuit is similar to those proposed by Nix and Laithwaite [1], Nasar and Boldea [2], Davis [5], and Groot and Heuvelman [6], but in Ref. [2,6] the stator leakage flux is considered.

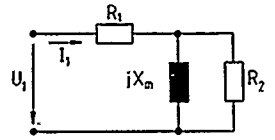


Fig. 1 - Equivalent circuit per phase at standstill, neglecting core losses and leakage flux

However, the expressions for TA parameters R_1, X_m and R_2 are different from author to author. These parameters may be determined by the following equations [2,4]:

$$R_1 = 2\rho_c\left(1/d_c^2\right)(D_e + D_i)N_1 \tag{1}$$

$$X_m = \left(6\mu_0\omega/\pi\right)\left(\tau_p D/\rho K_c g\right)N_1^2 \tag{2}$$

$$R_2 = \left(6\pi D/\tau_p p\right)\left(\rho/t\right)N_1^2 \tag{3}$$

The expression of standstill force, or thrust, derived from the equivalent circuit as well as from Maxwell's equations, is given by [1, 4, 7],

$$F = 18\pi^2 p D\left(\rho/t\right)\left(I_1^2/\omega\right)\left(z/\tau_p\right)^2\left[Q^2/\left(1+Q^2\right)\right]K \tag{4}$$

where Q is the Laithwaite's goodness factor [1] and K_c is the Carter's coefficient [8]. Q and K_c are expressed as,

$$Q = \left(\tau_p^2\mu_0\omega t\right)/\left(\pi^2\rho K_c g\right) \tag{4a}$$

$$K_c = \tau_s/\left[\tau_s - s^2/(s+5g)\right] \tag{4b}$$

The inclusion of K_c is important and necessary in the calculation of force [6,7].

The force correction factor K[1,7] is a function of pole number only, and becomes simply equal to 1 for even-pole TAs

3. EXPERIMENTAL RESULTS

Figs. 2, 3 show the comparison between measured results and the theoretical curves, obtained in the test of prototypes TAI and TAII (see Appendix). The degrees of error that result of this comparison are: 1 - Power factor - TAI: 9% including K_c and 10% not including K_c; TAII: 2% including K_c and 18% not including K_c. 2 - Standstill force - TAI: 12% including K_c and 50% not including K_c; TAII: 8% including K_c and 40% not including K_c. Consequently, the importance of K_c in the calculation of force is corroborated by these results. A comparison between experimental and theoretical values, including K_c, shows a good agreement, confirming the validity of equivalent circuit.

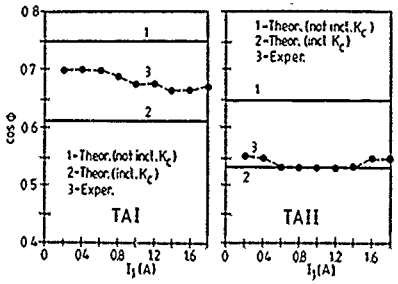


Fig. 2 - Power factor against r.m.s. phase current at standstill (50 Hz)

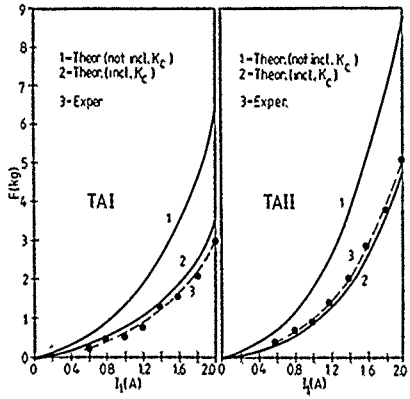


Fig. 3 - Standstill force against r.m s. phase current (50 Hz)

4. OPTIMISATION OF DESIGN

For the following imposed data: t=0.0015 m, g=0.0020 m, K_c = 1.5, Q = 1, ω=314 1/s, from (4a), τ_p=0.039 m [4]. Thus, the optimized design formulas becomes:

$$R_1 = 3.4.10^{-8}\left(1/d_c^2\right)\left(D_e + D_i\right)N_1 \qquad (5)$$

$$X_m = R_2 = 10^{-2}\left(D/p\right)N_1^2 \qquad (6)$$

$$F = 9.10^{-4}\left(D/p\right)\left(N_1 I_1\right)^2 K \qquad (7)$$

This optimum design criterion is based in fact that the maximum value of force per power input is maximum at standstill when Q=1 [1, 2, 4, 5].

5. CONCLUSIONS

From analytical equivalent circuit, whose validity was tested experimentally by means two prototypes, a simple design formulas provide a reliable method to the optimized design of TAs for standstill applications.

6 - REFERENCES

[1] - Nix, G. F., and Laithwaite, E.R.: "Linear induction motors for low speed and standstill application", Proc. IEE, June 1966, p. 1044.

[2] - Nasar, S.A., and Boldea, I: "Linear motion electric machines", Ed. John Wiley, 1976.

[3] - McLean, G.W.: "Review of recent progress in linear motors". Proc. IEE - B, Nov. 1988, p. 380.

[4] - Cabrita, C.P.: "Motor Linear de Indução. Análise Teórica, projecto e ensaio", Ph. D. Thesis, IST, Lisbon, 1988.

[5] - Davis, M.W.: "Development of concentric linear induction motor". Proc. IEEE - PAS, July 1972, p. 1506.

[6] - De Groot, D. J., and Heuvelman, C.J.: "Tubular linear induction motor for use as a servo actuator". Proc. IEE - B, July 1990, p. 273.

[7] - Cabrita, C.P.: "Força generalizada em Motores Lineares". Technical report, S.M.E.E.P./I.S.T., Lisbon, 1985.

[8] - Martin, J. C.: "Cálculo industrial de Máquinas Eléctricas". Ed. Danae, Barcelona, 1968.

APPENDIX. Details of experimental tubular actuators TAI and TAII:

	TAI	TAII		TAI	TAII
2p	4	3	D, mm	17.5	43.3
z_c	400	400	D_e,mm	60	90
N_1	1600	1200	D_i,mm	28	54
τ_p, mm	43.5	43.5	g, mm	4.50	4.25
s, mm	13	13	t, mm	1.5	1.3

COMPARATIVE STUDY OF SIMULATION METHODS OF INVERTER FED SYNCHRONOUS MOTOR

BERNARD DAVAT, RENE LE DOEUFF AND MOHAMMED ASSINI
Groupe de Recherche en Electrotechnique et Electronique de Nancy
ENSEM, INPL, 2 Av. de la Forêt de Haye, 54516 Vandoeuvre lès Nancy, France.

Département de Génie Electrique
ENSEM, B.P. 8118, Oasis, Casablanca, MAROC

Abstract - This paper investigates the use of two different simulation methods of converter fed electrical machines. The first one called global simulation method needs only the knowledge of the converter diagram, the different firing sequences and the machine parameters. The second one, based on a sequential analysis of the converter operating sequences uses an automatic generator of numerical simulation program of electro-mechanical process.

I. INTRODUCTION

An analytical study of power electronic system including static converter can be done when the different working sequences are known, and when their successions in time domain are predicted. Otherwise, numerical simulation tools become helpful mean to investigate the behaviour of such systems. These programs can be classified into two groups. In the first one, the simulation method discovers the functionning of the converter, step by step, from the only knowledge of the converter diagram and the semiconductors firing sequences. In the second one, the different working sequences are previously known in order to limit tests on succession of them. The two packages which are presented in this paper belong to these two groups: SACSO-Machine [1], an automatic program for the simulation of converter fed machine is a part of a family (SACSO, SCRIPT) still developped actualy; GASPE [2], a simulator, generates automatically, from the previous knowledge of the different operating sequences, a simulation program of any control-converter-machine assembly.

After the presention of these two packages, simulation results of the functionning of an inverter fed synchronous motor (Fig. 1) show the advantages and the drawbacks of employed simulation methods. These comparisons lead to point out the best using field of both packages.

II. SACSO-MACHINE PACKAGE

SACSO-Machine is composed of three main parts: a data unit, where is described the converter topology, the control of the semiconductors and the machine parameters, a simulation processor of the machine converter set and a result post processor. The semiconductors are considered as resistances whose values depend on if they are blocked or conducting. Each sequence of the converter is described by a state equation which depend on the topology and the states of the semi-conductors. Combining this equation with the electrical machine equations expressed in a Park reference frame leads to an equation which represents the functionning of the whole system

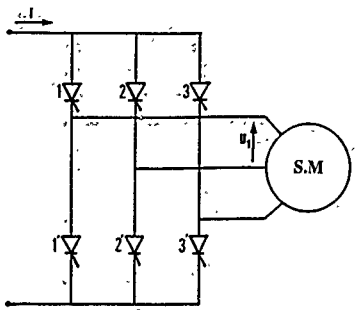


Fig. 1. Circuit diagram.

$$dX/dt = A \cdot X \qquad (1)$$

in which the X vector is composed of the converter state variables, of the currents in the machine coils (in the Park reference frame) and of a term representing the supplying sources of the machine-converter set. Equation (1) is solved in a form of a recurrent equation

$$X(t+T) = \mathrm{Exp}(A.T) \cdot X(t) \qquad (2)$$

In order to reduce the simulation time which essentially depends on the matrix exponential calculation, one can transform equation (1) in

$$dZ/dt = D \cdot Z \qquad (3)$$

D being a diagonal matrix, the exponential matrix is then obtained analytically from the exponential of the eigenvalues of A.

In order to accelerate the transient state, a computation of the steady state solution based on the Newton Raphson method is implemented in the SACSO-Machine package.

III. GASPE METHODOLOGY

GASPE generator has been designed in order to study power electronic systems which are generally composed of an association of a static converter, an electric machine and control circuits. Actually, it generates a simulator which is made of a main program body, a simulation management system and a block procedure to detect all occurences of events and updating variables (time, currents, semiconductor states...).

After analysing the system to be simulated, the first step consists to subdivide the system into blocks with respect to its function. For the considered problem there are a Motor Block for the synchronous motor (here a non salient one modelling with the Park

transformation), a Source Block to represent the current source which is assumed here to be constant and a Thyristor Control Block to generate the firing signals of the thyristors according to the rotor position.

In order to use GASPE methodology, a set of informations in a specific language has to be written. The Grammar File, which specifies the type of variables and the nature of events for each block.

For the simulation of current inverter fed synchronous motor, a fast convergence process is implemented in the simulator in order to reach rapidly the steady state.

IV. SIMULATION RESULTS

Simulation results of steady state operation of the inverter fed machine are presented in Fig. 2. The research of the steady state obtained from the two programs is shown in Fig. 3. The steady state is obtained after one period with GASPE due to the particular implemented method.

In SACSO-Machine package, the resolution method is based on the calculation of a matrix exponential. This method has the advantage to allow an analytical solution in the form of a recurrent equation where the calculation step does not depend on the time constants of the system. Nevertheless, this method is not really compatible with the study of mechanical transient operation, neither the modelling of a salient pole machine. However this simulation method allows to study unusual or unpredictable working sequences (Fig. 4).

In the simulator generated by GASPE, the functionning of the converter has to be defined previously. Then, unusual working sequences cannot be studied. However, this package takes into account salient pole machine (Fig. 4) and sophisticated control loops.

V. CONCLUSION

This paper investigates the use of two simulation methods to study the behaviour of synchronous machine supplied by a current source inverter. The advantages of the two methods have been point out. SACSO-Machine program becomes necessary when the working sequences are unusual or unpredictable, either in transient state or following a fault. Programs developped from the GASPE methodology do not use a so sophisticated resolution method and then allow to study salient pole machines or complicated control loops including mechanical transient functionning. However these two packages are complementary, the advantages of one make up for the drawbacks of the other.

REFERENCES

[1] - Davat (B.), Foch (H.) and Trannoy (B.), Analysis of a converter fed asynchronous machine by a global simulation method. Electric Machines and Electromechanics, Vol. 5, 1980, pp. 201-210.

[2] - Simonot (F.), Le Doeuff (R.), Haddad (S.) and Ramaromisa (V.), An Object Oriented System for the Automatic Generation of Simulation Programs in Power Electronics. International Conference on CAD, Beijing (China), August 1989.

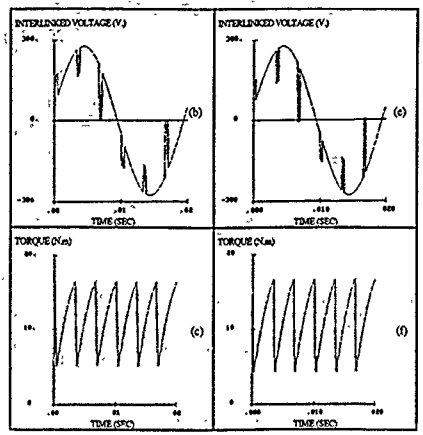


Fig. 2. Steady state operating ; SACSO-Machine results (left) and GASPE results (right).

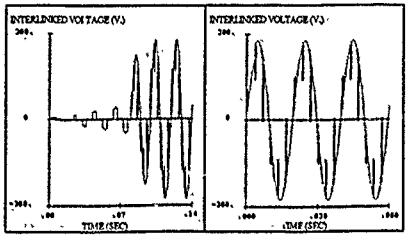


Fig. 3. Steady state research, SACSO-Machine (left) and GASPE (right).

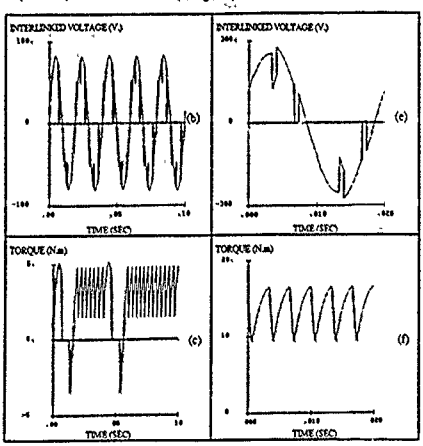


Fig. 4. Missing of a firing pulse of a thyristor with SACSO-Machine (left) and salient pole machine with GASTE (right).

DIGITAL CONTROL OF A PM SYNCHRONOUS ACTUATOR DRIVE SYSTEM WITH A GOOD POWER FACTOR

Y. Fu S. Robyns

Université Catholique de Louvain
Laboratoire d'Electrotechnique et d'Instrumentation
Bâtiment Maxwell - Place du Levant, 3
B-1348 Louvain-la-Neuve
Belgica

Abstract - This paper discuss, for PM synchronous actuators, the limits of application of the control method which keep the direct axis component of the armature current equal to zero, and describes two control algorithms which maintain the power factor as close as possible to unity. These algorithms have been evolved with a view to an easy hardware implementation.

The performances of the proposed control algorithms have been validated by digital simulation.

1. INTRODUCTION

For permanent magnet synchronous actuators, the control method consisting of keeping the direct axis component of the armature current equal to zero allows to use very simple control algorithms [1,2]. This method is the most commonly used, but can strongly deteriorate the power factor if applied to some type of buried magnet synchronous actuators.

Control methods which improve the power factor are more intricate because they need a non linear control law. However these control methods reduce the stator voltage allowing to use a lower rated voltage inverter. On the other hand, as modern permanent magnet actuators can reach tens of kW [3], to keep a good power factor allows to improve the efficiency.

This paper determines, in accordance with the rotor saliency, the limits of application of the control method which keep the direct axis component of the armature current equal to zero, and describes two control algorithms which maintain the power factor as close as possible to unity. These algorithms have been evolved with a view to an easy hardware implementation.

The performances of the proposed control algorithms have been validated by digital simulation.

2. ACTUATOR MODELLING

Permanent magnet synchronous actuators came in two configurations according to their rotor geometry (figure 1) [4] :

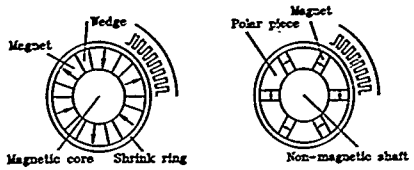
- PM synchronous actuators with surface mounted magnets which schematic representation is shown in figure 1.a have no significant saliency effects. Their saliency coefficient $\rho = L_q/L_d$ is nearly equal to unity.
- Buried PM synchronous actuators which schematic representation is shown in figure 1.b have a saliency coefficient which can reach values as high as 5.

For a two pole actuator without damper windings, the Park equations reduce to :

$$\begin{vmatrix} V_d \\ V_q \end{vmatrix} = \begin{vmatrix} R_s+L_d\,s & -\vartheta_m L_q \\ \vartheta_m L_d & R_s+L_q\,s \end{vmatrix} \begin{vmatrix} i_d \\ i_q \end{vmatrix} + \begin{vmatrix} 0 \\ K_T\,\vartheta_m \end{vmatrix} \quad (1)$$

$$T_{em} = K_T\, i_q + (L_d-L_q) i_d\, i_q$$

In these equations
- L_d and L_q are the self inductances of the d and q armature equivalent windings
- R_s is the resistance of the armature dq windings
- K_T the torque constant
- ϑ_m the angular position of the rotor and $\dot{\vartheta}_m$ its velocity
- s the Laplace operator.



a b

Fig. 1. Schematic representation of PM synchronous actuators.

3. THEORETICAL STUDY

The control strategy and the rotor geometry affect the performances of the system and the size of the inverter feeding the actuator. Three control strategies are investigated : the first one consists of keeping $i_d=0$, the second one of keeping $\cos\varphi=1$ and the third one of keeping the RMS value of the stator voltage proportional to the rotor speed. Some other theoretical considerations which don't appear in this study are given in reference [5].

In order to simplify the theoretical study, the currents i_d and i_q are replaced by the variables I_s and β according to equations (2).

$$I_s = \sqrt{i_d^2 + i_q^2} \quad (2.a)$$

$$\beta = \text{arctg}\,(-\frac{i_d}{i_q}) \quad (2.b)$$

I_s is the RMS value of the stator current (multiplied by $\sqrt{3/2}$ for a three phase actuator) and β is its leading angle with respect to the rotor q axis.

With these variables, equations (1) become in steady-state :

$$\begin{vmatrix} V_d \\ V_q \end{vmatrix} = \begin{vmatrix} R_s & -\vartheta_m L_q \\ \vartheta_m L_d & R_s \end{vmatrix} \begin{vmatrix} -I_s \sin\beta \\ I_s \cos\beta \end{vmatrix} + \begin{vmatrix} 0 \\ K_T\,\vartheta_m \end{vmatrix} \quad (3)$$

From equations (3), the vector diagram shown in figure 2 is obtained.

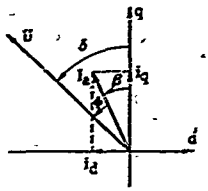


Fig. 2. Vector diagram.

The electromagnetic torque becomes

$$T_{em} = I_a \left[K_T \cos\beta + \frac{1}{2}(\rho-1) L_d I_a \sin 2\beta\right] \qquad (4)$$

where, $\rho = \frac{L_q}{L_d}$ is the saliency coefficient.

The motor power factor is

$$\cos\varphi = \cos(\delta-\beta)$$

where

$$\operatorname{tg}\delta = -\frac{V_d}{V_q} = \frac{R_s I_a \sin\beta + \partial_m L_q I_a \cos\beta}{R_s I_a \cos\beta - \partial_m L_d I_a \sin\beta + K_T \partial_m} \qquad (5)$$

These equations are normalized as follows

$$I_a^N = \frac{L_d I_a}{K_T} \qquad T^N = \frac{T_{em} L_d}{K_T^2}$$

By neglecting the voltage drop on the armature resistance R_s in comparison with the back emf, we can rewrite (4) and (5) in function of the normalized variables I_a^N and T^N :

$$T^N = I_a^N \left[\cos\beta + \frac{1}{2}(\rho-1) I_a^N \sin 2\beta\right] \qquad (6)$$

$$\operatorname{tg}\delta = \frac{\rho I_a^N \cos\beta}{1 - I_a^N \sin\beta} \qquad (7)$$

* Control with $i_d = 0$

In this control method the direct axis component of the stator current i_d doesn't exist. The relation between the torque and the current is linear and this allows to evolve very simple control algorithm [1,2].

Figure 3 shows that the power factor decreases strongly with the load for actuator with a saliency coefficient greater than the unity.

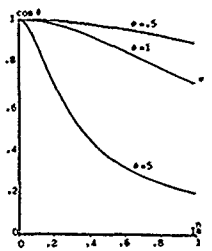


Fig. 3. Power factor with the i_d=0 control method.

* Control with $\cos\varphi = 1$

In order to keep the power factor always equal to unity, it is necessary to achieve the following relation between I_a^N and β.

$$I_a^N = \frac{\sin\beta}{\sin^2\beta + \rho\cos^2\beta} \qquad (8)$$

One drawback of this method is that the relation between the torque and the current is a non linear one, as shown in figure 4.a.

This figure shows also that the torque decreases beyond a certain value of I_a^N. Nevertheless a lot of actuators have a rated current lower than that value of I_a^N.

However it is possible to maintain the torque-current relation roughly linear by keeping β constant beyond a critical current judiciously chosen. This is achieved with a small deterioration of the power factor, but it remains better than with the control keeping i_d=0 for actuator with $\rho > 1$.

Figure 4.b shows the power factor and the torque characteristics when β is kept constant beyond a critical value of the current I_a^N.

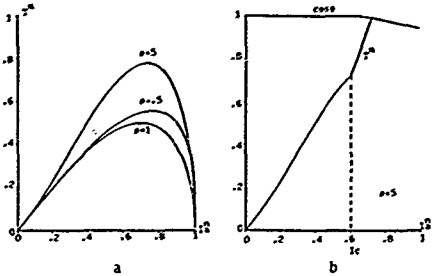


Fig. 4. Torque and power factor with the cosφ=1 control method.

* Control with RMS value of the stator voltage proportional to the speed

To keep the RMS value of the stator voltage constant and equal to the back emf, it is necessary to achieve relation (9).

$$I_a^N = \frac{2\sin\beta}{\sin^2\beta + \rho^2\cos^2\beta} \qquad (9)$$

Figure 5 shows the power factor and the torque characteristics for this control method.

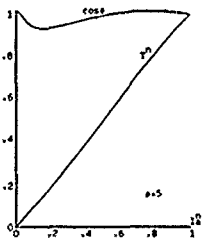


Fig. 5. Torque and power factor of the control with RMS value of the stator voltage proportional to the speed

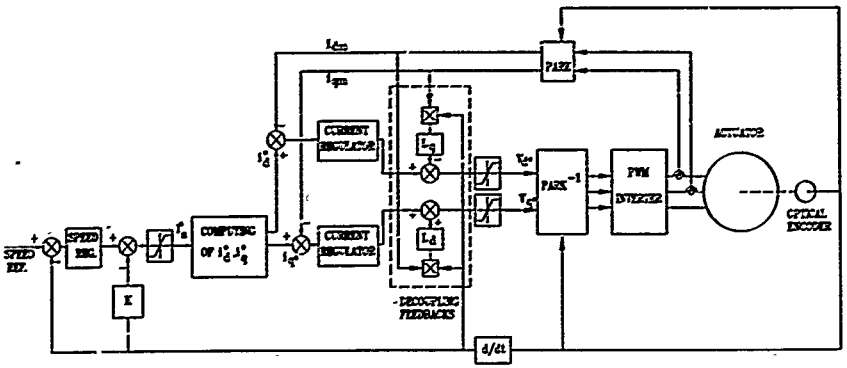


Fig. 6. General overview of the proposed controller.

The torque-current relation is practically linear whereas the power factor remains acceptable.

4. SYNTHESIS OF THE DIGITAL CONTROLLER

The control with cosϕ=1 and the control keeping the RMS value of the stator voltage proportional to the speed can both be achieved by the controller shown in figure 6.

The difference between the two control methods is in the content of the block called "computing of i_d^*, i_q^*".

For the control with cosϕ=1, this block determines if I_s^* is smaller or bigger than the chosen critical value I_c^* of I_s^*. Then it computes i_d^* and i_q^*. The algorithm is the following.

If $I_s^* \leq I_c$, β is the solution of (8) :

$$\sin\beta = \frac{1 - \sqrt{1 - 4\, I_s^{N}(1-\rho) I_s^{N}\, \rho}}{2\, I_s^{N}(1-\rho)} \qquad (10)$$

If $I_s^* > I_c$, β remains constant and corresponds to the value computed from (10) when $I_s^* = I_c$.

For the control with RMS value of the stator voltage proportional to the speed β is determined by solving (9). This yields (11).

$$\sin\beta = \frac{1 - \sqrt{1 - I_s^{N}(1-\rho^2) I_s^{N}\, \rho^2}}{I_s^{N}(1-\rho^2)} \qquad (11)$$

In both cases, i_d^* and i_q^* are computed from I_s^* and β.

$$i_d^* = -I_s^* \sin\beta$$
$$i_q^* = I_s^* \cos\beta \qquad (12)$$

To reduce the computing time, it is possible to tabulate the values of i_d^* and i_q^* as a function of I_s^* according to equation (10) or (11).

i_d^* and i_q^* are the references values for two current regulators which are here two PI regulators. The synthesis of these two regulators is easy if a decoupling between the two axes is realized through an appropriate state feedback. The values of V_d^* and V_q^* are obtained by adding to the outputs of the current regulators supplementary terms computed by the decoupling feedback unit.

Furthermore a feedback on the speed regulator s output of a well chosen term proportional to the speed $k\dot{\delta}$ improves strongly the insensivity of the system against load variations.

The Park and inverse Park transform involved in the control algorithm are performed without lost of time by using tabulated values of $y=x_1 \cos(x_2)$ [1,6].

Another way to control the actuator currents consists in controlling each phase current by a local feedback loop acting on the corresponding leg of the inverter feeding the motor. The problems related to these local feedback loops have been investigated in reference [7].

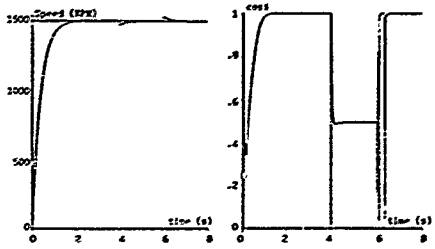
The two proposed control methods are more intricate than the control with $i_d=0$. Indeed the control with $i_d=0$ can be achieved by using only one PI regulator for the speed and one proportional amplifier for the control of i_q, since the condition $i_d=0$ may be ensured by an appropriate decoupling feedback [2]. Furthermore if the electrical time constant may be neglected, the proportional amplifier is no more necessary and it is also possible to avoid current measurements [1].

5. PERFORMANCES OF THE SYSTEM

Simulations allow to compare the performances of the three control methods discussed in this paper. The parameters of the simulated actuator are the following ones.

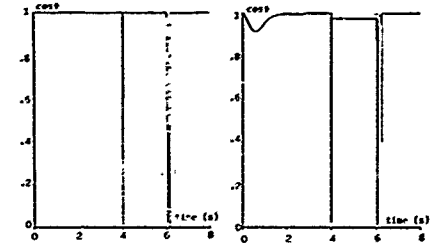
R_s = 1 Ω	K_T = 2,06 V.S/rad
L_d = 0,038 H	J = 0,08735 kgm²
L_q = 0,197 H	K_f = 0,00342 Nms

Its saliency coefficient ρ=5 is high. Figure 7 shows the response of the actuator to a step in the speed reference and to the application of a positive step of torque (nearly equal to 70% of the rated torque) at t=4s and a negative one at t=6s. If the speed's performances are similar for the three control algorithms, figures 7.c and 7.d show the improvement of the power factor achieved by the proposed control algorithms.



a. Actuator speed response

b. Control with i_d=0



c. Control with cos$=1

d. Control with RMS value of the stator voltage proportional to the speed

Fig. 7. Simulation of the three control methods.

6. CONCLUSION

In this paper, two control algorithms which allow to keep the power factor as close as possible to unity have been proposed. These two control algorithms, quasi equivalent, are a little more intricate than the one which achieve the commonly used control with i_d=0. But their hardware implementation can be done without important additional difficulties.

The use of control algorithms which maintain the power factor close of unity is justified only for actuators with a saliency coefficient significatively greater than unity, as buried PM synchronous actuators. The i_d=0 control method remains the most interesting one for actuators without saliency.

7. REFERENCES

[1] H. Buyse, Th. Canon, J.Ph. Conard, F. Labrique, P. Sente
Digital field oriented control of a PM synchronous actuator without current sensors
Proceedings of the third EPE Conference, Aachen, 9-12 October 1989, pp. 1067-1072.

[2] H. Buyse, F. Labrique, B. Robyns, P. Sente
Digital field oriented control of a PM synchronous actuator using a simplified strategy for controlling the Park components of the stator currents.
IMACS TCI'90, International Conference on Modelling and Simulation of Electrical Machine and Static Converters, Nancy, 19-21 September 1990, pp. 37-41.

[3] W. Leonhard
Adjustable speed AC drives
Transactions of the IEEE, vol.76, n°4, April 1988, pp. 455-470.

[4] M. Lajoie-Mazenc
Utilisation des aimants permanents dans les machines à commutation électronique
Journées d'Etudes INPL-SEE, Machines électriques et techniques de pointe, Nancy, 21-22 septembre 1983.

[5] Y. Takeda, S. Morimoto, T. Hirasa, K. Fuchi
Most suitable control method for permanent magnet synchronous motors
Proceedings of ICEM 88, Pisa, 12-14 September 1988, pp. 53-58.

[6] P. Sente, H. Buyse
Modulateur de largeur d'impulsion pour onduleur à commande numérique vectorielle
6e Colloque sur le Positionnement incrémental par entraînement électrique, Lausanne, 5-6 July 1990, pp. 101-112.

[7] M. Lajoie-Mazenc, C. Villanueva, J. Hector
Study and implementation of hysteresis controlled inverter on a permanent magnet synchronous machine
IEEE Transactions on Industry Applications, vol.IA-21, n°2, March/April 1985, pp. 408-413.

A ROBUST POSITION CONTROLLER FOR A DC DRIVE USING SLIDING MODE AND AN ACCELERATION OBSERVER

Dente, J. A.; Maia, J. H.

Secção de Máquinas Eléctricas e Electrónica de Potência Instituto Superior Técnico IST / INIC
Av. Rovisco Pais 1096 Lisboa Codex Portugal.

Abstract - A simple way for synthesizing a robust position controller for a DC drive is presented. The applied method uses the input output linearization method to achieve a more adequate representation of the system. The robustness is assured by the use of the sliding mode operation.

I. INTRODUCTION

In this paper is presented a principle of synthesizing a robust position controller for a DC drive using sliding mode. This method enables to define in an easy way the commutation law that assures required performances and the robustness of the system. This commutation law supposes that a correct acquisition of the position, speed and acceleration is possible. Concerning implementation, the acceleration value may put some practical problems, because a large frequency bandwidth transducer is not available and noise level do not enable us to compute the speed derivative. To overcome this drawback, an acceleration observer is used.

Experimental results show the good performances achieved with the proposed method for position regulation and for position tracking.

II. THEORETICAL STUDY

The system is based on a separated excited DC motor and is represented by the equations (1), where T(t) is the load torque.

$$\begin{cases} \dfrac{di}{dt} = -\dfrac{R}{L} i - \dfrac{K_\phi}{L} \omega + \dfrac{u}{L} \\ \dfrac{d\omega}{dt} = \dfrac{K_\phi}{J} i - \dfrac{1}{J} T(t) \\ \dfrac{d\theta}{dt} = \omega \end{cases} \tag{1}$$

This model results from the application of circuit theory laws and is deliberately simplified. In practice, the magnetic circuit is not linear and there exists parameter variations.

The model described by equations (1) is simple. However, it is not the more adequate to synthesize a robust position controller. For this purpose, it is useful to chose another state variables: the position θ, speed ω and acceleration γ. It is important to refer that the acceleration reports a "large quantity of information", because this variable contains information about the current and also about the load torque. This new model is obtained by the dipheomorfism represented by equations (2), which could be obtained in an systematic way like refered in [1] and [2].

$$\begin{cases} \theta = \theta \\ \omega = \omega \\ \gamma = \dfrac{1}{J}\left(K_\phi i - T(t)\right) \end{cases} \tag{2}$$

Using the error values (3), the system will be represented by equations (4).

$$e_\theta = \theta_{ref} - \theta \qquad e_\omega = \dot{\theta}_{ref} - \omega \qquad e_\gamma = \ddot{\theta}_{ref} - \gamma \tag{3}$$

With the result stated by (4) and using as applied voltage the value given by (5) where the commutation law S is a linear function of the errors, a robust tracking position control is achieved. This result is in according with another one theoretically presented in reference [3].

$$\begin{cases} \dfrac{de_\theta}{dt} = e_\omega \\ \dfrac{de_\omega}{dt} = e_\gamma \\ \dfrac{de_\gamma}{dt} = A_\gamma(e_\omega\, e_\gamma\, t) - B_\gamma u \end{cases} \tag{4}$$

$$u = V \operatorname{sgn}(S) \; ; \; S = K_\theta e_\theta + K_\omega e_\omega + K_\gamma e_\gamma \tag{5}$$

The dynamic of the system is now represented by equation (6), if the sliding mode control is effective.

$$e_\gamma + 2\xi\omega_0 e_\omega + \omega_0^2 e_\theta = 0 \tag{6}$$

III. IMPLEMENTATION

As indicated by equations (3) and (4) and concerning implementation, it is necessary to measure the acceleration value. This may put some practical problems, because a large frequency bandwidth transducer is not available and noise level do not enable us to compute the speed derivative. To overcome this drawback, an observer for the external input (torque T and also parameter variations) is used. This signal is combined with the true measure of the current to achieve an approximate value of the acceleration.

The model of this observer, also used in reference [5], is as follow:

$$\begin{cases} \dot{\delta} = -g\,\delta + g\,K_{\phi n}\, i + g^2 J\,\omega \\ \hat{T}_{ext} = \delta - g J\,\omega \\ \hat{\gamma} = \dfrac{1}{J}\left(K_{\phi n}\, i - \hat{T}_{ext}\right) \end{cases} \tag{7}$$

where g is a theoretically free chosen parameter.

IV. EXPERIMENTAL RESULTS

Experimental results were obtained using a 1.5 kW separated excited DC motor supplied by a four quadrants GTO DC/DC converter The output variable θ is measure by a shaft incremental encoder. To measure the speed it was also used this encoder and a dedicate electronic circuit [6]. The perturbation torque T is generated by a gearbox connected pendulum, as the one presented on figure 1.

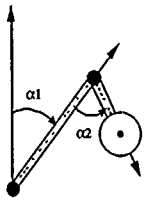


Fig. 1 - Pendulum used as motor load.

The error signal S of equation (5) with the current limitation is formed using a simple electronic circuit like the one presented on

figure 2 by a block diagram. However, one must note that the solution presented in this figure is only enough for regulator applications. Indeed, when the reference changes, i.e. for tracking applications, it is necessary to consider the first and second derivatives of the reference, as it is shown by the experimental results in figure 3

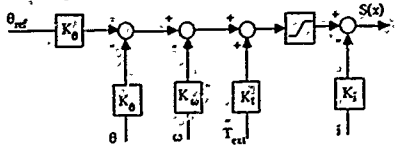
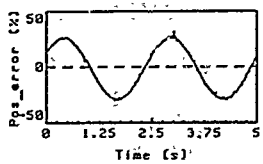
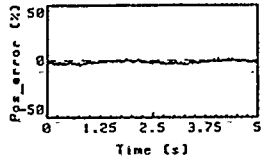


Fig. 2 - Error calculating block diagram for a regulator application.



a) first and second derivatives of the reference are not considered.



b) first and second derivatives of the reference are considered

Fig. 3 - Experimental recording of the position error when a sinusoidal reference is applied [$\theta_{ref} = 0.3 \sin(2\pi v 2.5)$].

There are minor motives that cause the error presented in figure 3-b), for instance the frequency limitation in the DC/DC converter, and the approximate mesure of the speed However, this error is mainly due to low pass filter characteristics of the acceleration observer In figure 4 this fact is underlined by the error evolution when a sudden action at the pendulum is made.

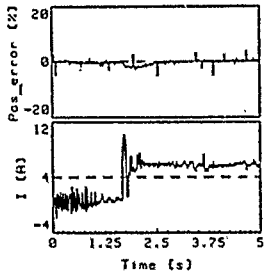


Fig. 4 - Position error and motor current evolution when a sudden load is applied to the motor shaft.

There is another important source of error for the system. This happens when the compatibility of the reference variation with the physics of the motor is not assured. In figure 5 is presented the error evolution when the described situation is created by a step input signal. Also in this figure one can see the current limiting action.

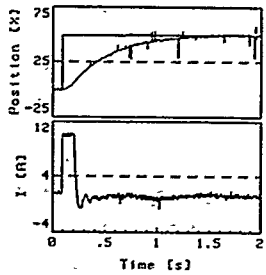


Fig. 5 - Position and current evolution when a step reference is applied.

In figure 6 the error is presented when a 50% variation on field current is made. The result shows the insensitivity to K_θ parameter, wich variation is compensed by a large current value.

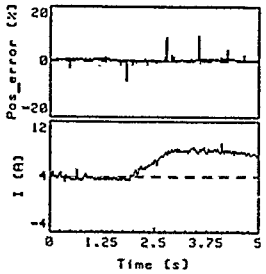


Fig. 6 - Error and current evolution when a field current variation is generated.

V. CONCLUSIONS

Theoretical study and experimental results show that the proposed principle for synthesizing a position controller is easily applied and gives good performance for regulation or tracking applications. Insensitivity concerning parameters and supply voltages changes and also perturbation due to load variation is achieved by the application of the sliding mode control with a suitable choice of the commutation law.

Restrictions in the performances are only imposed by the physics of the system and the availability of direct mesures of the state variables. At a secondary level restrictions are also imposed by the execution of the control electronics.

VI. REFERENCES

[1] - Campion G.; Bastin G.: "Indirect Adaptive State Feedback Control of Lineary Parametrized non Linear Systems", Technical Report AP 89.09 Université Catholique de Louvain, Lab. d'Automatique, Dynamique et Analyse des Systèmes. May 1989

[2] - Dente, A.; Garrido, M. S.; Maia, J. H.: "A robust tracking speed control of a permanent magnet synchronous motor", IMACS'TC1, Nancy, pp. 21-24, 1990.

[3] - Utkin, V. I.: "Discontinuous Control Systems: state of art in theory and applications", IFAC-Munich, pp. 75-94, 1987

[4] - Ohishi, K.; Nakao, M.; Ohnishi, K.; Miyachi, K.: "Microprocessor-controlled DC Motor for Load-insensitive Position Servo System", IEEE Trans. on IE, vol. IE-34, nº1, pp. 44-49, 1987.

[5] - Dote, Yasuhiko: "Application of modern control techniques to Motor Control", Proceedings of the IEEE, vol. 76, nº4, pp. 438-454, 1988.

[6] - Bonert, R. "Digital tachometer with fast dynamic response implemented by a microprocessor", IEEE Trans. on IE, vol. IE - 34, nº1, pp. 11-18, 1983.

STABILITY ANALYSIS OF A DECOUPLING STATE FEEDBACK BASED FOC OF A PM SYNCHRONOUS ACTUATOR

B. Robyns F. Labrique H. Buyse

Université Catholique de Louvain
Laboratoire d'Electrotechnique et d'Instrumentation
Bâtiment Maxwell - Place du Levant, 3
B-1348 Louvain-la-Neuve
Belgium

Abstract - This paper deals with the stability analysis of an electrical drive system based on a PM synchronous actuator with field oriented control when field orientation is performed through a decoupling state feedback. It is shown that the use in the decoupling state feedback of predicted values of the actuator currents instead of measured ones improves substantially the stability of the system.

1. INTRODUCTION

This paper deals with the stability analysis of an electrical drive system based on a PM synchronous actuator with field oriented control when field orientation is performed in open loop by using an appropriate decoupling state feedback [1,2].

As field orientation is performed in open loop, its effectiveness depends on the uncertainties on the actuator's parameters and on the errors on the measured or computed values of the speed and the currents.

The aim of this paper is to investigate the effect of these uncertainties on the stability of the system. The study is performed by considering surface mounted PM actuators which present negligible electrical time constants.

The main result arising from the study is that the stability of the system is substantially improved by using for the computation of the decoupling feedback predicted values of the currents instead of measured ones.

2. ACTUATOR MODELLING

The Park equations of a PM synchronous actuator without damper windings reduce to

$$\begin{vmatrix} U_d \\ U_q \end{vmatrix} = \begin{vmatrix} R_a+L_d s & -P\vartheta_m L_q \\ P\vartheta_m L_d & R_a+L_q s \end{vmatrix} \begin{vmatrix} i_d \\ i_q \end{vmatrix} + \begin{vmatrix} 0 \\ K_T \vartheta_m \end{vmatrix} \quad (1)$$

$$T_{em} = K_T i_q + (L_d-L_q) i_d i_q \quad (2)$$

In these equations
- L_d and L_q are the self inductances of the d and q armature equivalent windings
- R_a is the resistance of an armature winding
- K_T the torque-constant
- θ_m the angular position of the rotor and ϑ_m its velocity
- 2P the number of poles
- s the Laplace operator.

The corresponding block diagram is given in full line in figure 1. In this figure T_r is the load torque and J is the inertia of the actuator's rotor and of the mechanical load.

3. CONTROL STRATEGY

In the dq frame, field orientation is achieved when i_d is equal to zero. This can be performed in open loop by introducing an appropriate state feedback (shown in broken line in figure 1). According to this state

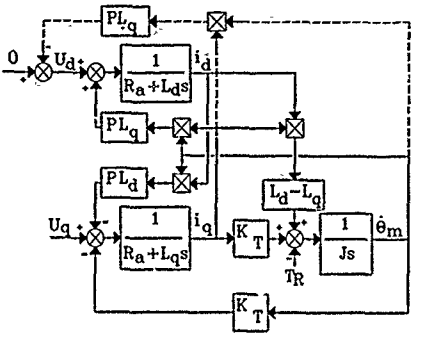


Fig. 1. Block diagram of a PM synchronous actuator without dampers.

feedback the d axis voltage must be equal to :

$$U_d = -P \vartheta_m L_q i_q \quad (3)$$

When the d axis current is equal to zero, the block diagram of the q axis becomes similar to that of a DC machine and the speed can be controlled by using a PI controller which generates the q axis voltage.

The above mentioned strategy leads normally to a digital controller. An example of implementation organised around two 8 bits MCS 8051 microcontrollers is described in reference [1].

4. STABILITY ANALYSIS

With the selected control strategy, the value of the d axis current depends directly on the precision in the generation of the voltage U_d according to equation (3).

In order to investigate the effects on the system stability of an error on the value of U_d, it is assumed that this voltage is determined by using an estimated value of L_q, L_q^*, affected by an error ΔL_q.

$$L_q^* = L_q + \Delta L_q \quad (4)$$

In this case, if the discretisation inherent to digital control and the small electrical time constants of the dq axes are neglected, one gets the block diagram of figure 2 for the q axis.

It can be seen in this figure that the error on L_q introduces a positive feedback loop if $\Delta L_q i_q > 0$. As this feedback depends of the square of the rotor speed, the tendency of the system to be unstable increase rapidly with the speed.

By linearising about a steady state working point the equations corresponding to the block diagram of figure 2, one gets the following equation :

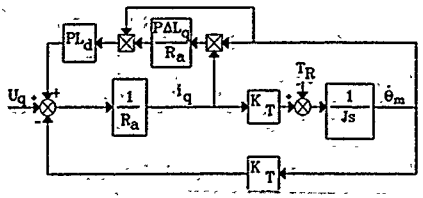


Fig. 2. Block diagram of the q axis with an unperfect decoupling.

$$\delta\theta_m = \frac{A}{1+CS}\,\delta U_q - \frac{B}{1+CS}\,\delta T_r \qquad (5)$$

with

$$A = \frac{1}{K_T - \frac{L_d}{R_a}\Delta L_q\, p^2\, 2\theta_{mo}\, i_{qo}}$$

$$B = \frac{R_a\left(1 - \frac{L_d}{R_a^2}\Delta L_q\, p^2\, \theta_{mo}^2\right)}{K_T\left(K_T - \frac{L_d}{R_a}\Delta L_q\, p^2\, 2\theta_{mo}\, i_{qo}\right)} \qquad (6)$$

$$C = \frac{JR_a\left(1 - \frac{L_d}{R_a^2}\Delta L_q\, p^2\, \theta_{mo}^2\right)}{K_T\left(K_T - \frac{L_d}{R_a}\Delta L_q\, p^2\, 2\theta_{mo}\, i_{qo}\right)}$$

In this equation, δU_q, δi_q, $\delta\theta_m$ and δT_r are the variations of the different variables about their steady state values U_{qo}, i_{qo}, θ_{mo} and T_{ro}.

The system shown in figure 2 is intrinsically stable if the real part of the pole of the transfer function (5) is negative. This implies the following conditions on ΔL_q :

$$\Delta L_q < \frac{1}{\frac{L_d}{R_a^2}\, p^2\, \theta_{mo}^2} \qquad (7.a)$$

and

$$\Delta L_q < \frac{K_T}{\frac{L_d}{R_a}\, p^2\, 2\theta_{mo}\, i_{qo}} \quad \text{if } \theta_{mo}\, i_{qo} > 0 \qquad (7.b)$$

$$\Delta L_q > \frac{K_T}{\frac{L_d}{R_a}\, p^2\, 2\theta_{mo}\, i_{qo}} \quad \text{if } \theta_{mo}\, i_{qo} < 0 \qquad (7.c)$$

Condition (7.a) which is independant of the current i_q is the most constraining one as it depends from the inverse of the speed's square.

In paragraph 3 it has been mentioned that the speed control may be achieved with a PI controller. The design of this controller has been described in reference [1]. The application of the Routh Hurwitz criterion to the closed loop system yields the following conditions for the stability of the system.

$$\Delta L_q < \frac{1}{\frac{L_d}{R_a^2}\, p^2\, \theta_{mo}^2} \qquad (8.a)$$

and

$$\Delta L_q < \frac{K_T + G}{\frac{L_d}{R_a}\, p^2\, 2\theta_{mo}\, i_{qo}} \quad \text{if } \theta_{mo}\, i_{qo} > 0 \qquad (8.b)$$

$$\Delta L_q > \frac{K_T + G}{\frac{L_d}{R_a}\, p^2\, 2\theta_{mo}\, i_{qo}} \quad \text{if } \theta_{mo}\, i_{qo} < 0 \qquad (8.c)$$

In equations (8) G is the gain of the PI controllers. One must notice that condition (8.a) is identical to condition (7.a). Conditions (8.b) and (8.c) are less restrictive than (7.b) and (7.c).

Figure 3 shows the stability limits according to equations (8.a), (8.b) and (8.c) for the system given in appendix. It can be seen that at high speed (6000 RPM), an error of 1.76% in excess on L_q is sufficient to cause instability.

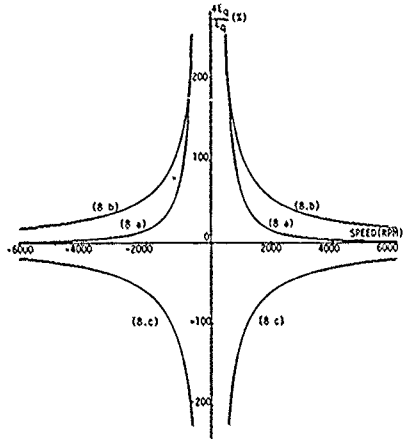
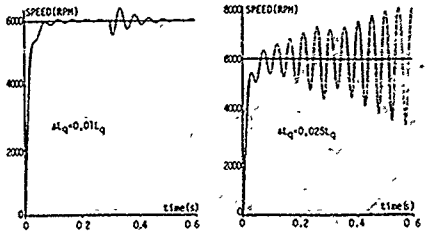


Fig. 3. Stability limit in the ($\Delta L_q/L_q$, speed) plane.

Simulations of the system have allowed to confirm the above mentionned theoretical results. In these simulations, the sampling period of the digital controller has been taken small enough (0,5ms) in order to reduce its influence on the behaviour of the system since it has been neglected in the theoretical analysis. Figures 4.a and 4.b show the response to a step in the speed reference from 0 to 6000 RPM and to the sudden application of a load torque equal to the rated torque at t=0,3s.

In figure 4.a, the decoupling feedback is computed by considering $L_q = 1{,}01 L_q$; at 6000 RPM the system remains stable even if one may notice some oscillations during the transient. In figure 4.b, the decoupling feedback is computed by considering $L_q = 1{,}025 L_q$ and according to the theoretical analysis the system becomes unstable.

It is worthy of note that the condition (7.a), (or (8.a)), is corresponding to an electrical instability. Indeed, if one assumes that the speed is constant one gets from figure 2 :



a b

Fig. 4. Simulation of the dynamic behaviour of the system using a measured value of i_q.

$$\delta i_q = \frac{\frac{1}{R_s}}{1 - \frac{L_d}{R_s^2} \Delta L_q \, p^2 \, \delta_{mo}^2} \delta U_q \qquad (9)$$

The denominator of (9) is positive when condition (7.a) is met.

5. DECOUPLING BASED ON AN ESTIMATED VALUE OF i_q

By replacing in the computation of U_d the measured value of i_q by an estimed value $\hat{i}_q$ [1,2]

$$\hat{i}_q = \frac{U_q - K_f \, \delta_m}{R_s} \qquad (10)$$

one gets, if the speed is assumed to be constant :

$$\delta i_q = \frac{p^2 \, \delta_{mo}^2 \frac{L_d L_q^*}{R_s} + R_s}{R_s^2 + p^2 \, \delta_{mo}^2 \, L_q \, L_d} \delta U_q \qquad (11)$$

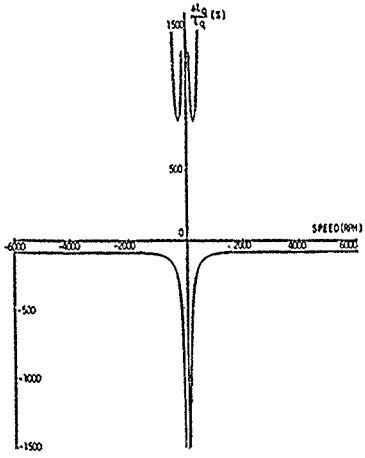


Fig. 5. Stability limit in the ($\Delta L_q/L_q$, speed) plane.

Equation (11) doesn't introduce any limitation on the maximum error on L_q.

If the mechanical part of the system is taken into account, the curves corresponding to the limit of stability in the ($\Delta L_q/L_q$, speed) plane become those of figure 5. It can be easily seen that the system stability is strongly increased. This is clearly apparent in figure 6. Indeed, this figure shows the response of the system to a step in the reference speed and to the sudden application of a load torque equal to the rated torque at t=0,3s, by assuming an error of 50% on the estimated value of L_q ($\Delta L_q = 0,5 L_q$). It can be seen that the system remains stable.

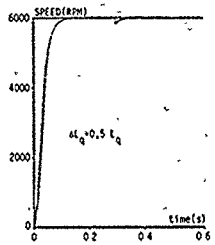


Fig. 6. Simulation of the dynamic behaviour of the system using an estimated value of i_q.

6. CONCLUSION

In this paper, one has investigated the stability of an electrical drive system based on a PM synchronous actuator with field oriented control if field orientation is performed through a decoupling state feedback. It is shown that it is necessary to use, in this feedback, a predicted value of the q axis current instead of a measured one in order to avoid a to high sensitivity of the system to modelling uncertainties as concern its limit of stability. The theoretical results have been validated by digital simulation.

7. REFERENCES

[1] H. Buyse, Th. Canon, J.Ph. Conard, F. Labrique, P. Sente
Digital field oriented control of a PM synchronous actuator without current sensors
Proceedings of the third EPE Conference, Aachen, 9-12 October 1989, pp. 1067-1072.

[2] H. Buyse, F. Labrique, B. Robyns, P. Sente
Digital field oriented control of a PM synchronous actuator using a simplified strategy for controlling the Park components of the stator currents
IMACS TC1'90, International Conference on Modelling and Simulation of Electrical Machine and Static Converters, Nancy, 19-21 September 1990, pp. 37-41.

APPENDIX

Parameters of the simulated actuator :

Rated power : 2 kW

Rated speed : 6000 RPM

P = 3

$R_s = 0,55\Omega$	$K_f = 0,297$ Nm/A
$L_d = 2,2$ mH	$J = 6.10^{-4}$ kgm^2
$L_q = 2,2$ mH	$K = 9,5.10^{-5}$ Nms/rad

AN ON-LINE FFT IMPLEMENTATION FOR A PARALLEL COMPUTER SIMULATION OF A VSI-IM RAIL TRACTION DRIVE

R. JOHN HILL and FENGTAI HUANG

School of Electrical Engineering, University of Bath,
Claverton Down, Bath BA2 7AY, England

Abstract - An on-line interactive FFT calculation of the harmonic spectra of the voltage and current variables in a VSI-IM rail traction drive simulation is presented. The implementation is on a parallel computer and is user-driven by a 'freeze and display' command. Simulation results are give for the spectra of the line and motor current waveforms during synchronous PWM operation of the drive.

I. INTRODUCTION

A major simulation task in a VSI-IM rail traction drive is to determine the harmonic performance as the VSI modulation strategy is changed. Consequently, most industrial simulators based on PCs or workstations incorporate a harmonic analysis facility to continuously display the spectra of the main electrical and mechanical variables throughout the traction duty cycle. The problem addressed in this paper is to provide a flexible selection facility, under interactive user control, for the extraction of harmonic information during the simulation of a such a traction drive.

In the paper, a parallel-computer based VSI-IM rail traction drive simulator is outlined. The software structure to enable on-line window definition and extraction of FFT-derived spectra in real time is then described. The procedure is carried out under user-interactive control via a freeze and display facility. Sample results for the harmonic spectra of input and motor currents during synchronous PWM operation in the acceleration period of the drive are then presented.

II. CONVERTER-MOTOR SYSTEM MODELLING

The VSI-IM rail traction drive modelled is illustrated in Figure 1. It is operated under practical conditions with a three-part duty cycle. From standstill to motor base speed, the IM is driven at constant flux; this is followed by constant power motion, with a final region of reduced power up to maximum speed. To achieve this duty cycle, the inverter is operated by PWM to the base speed (asynchronous PWM at startup, followed by synchronous PWM), and thence by quasi-square-wave modulation. In the PWM region, various modulation schemes are available which have the effect of reducing or minimising harmonic distortion of the supply current and motor voltage waveforms. However, these involve sudden change of the modulation ratio and/or index at certain speeds, which leads to poor harmonic performance at the changeover point.

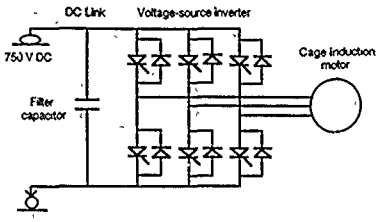


Fig. 1. VSI-IM rail traction drive.

The converter-motor system model, fully described in reference [1], simulates the filter, VSI and IM. The IM is modelled in the time domain by setting up and solving the motor differential equations. Transformation to a fixed D-Q axis model is then achieved giving the model equation in terms of the Laplace operator p as

$$[V] = \{[R] + \theta'[G] + [L]p\}[I] \qquad (1)$$

where [V], [I] are the direct and quadrature variable matrices, [R], [L], [G] are the elements of the IM rotor and stator equivalent circuits, and θ' is the rotor speed. The input DC filter is modelled as part of the IM impedance matrix, thus including the system input current and DC link voltage as additional state variables. The solution of Equation (1) is obtained by the method of eigenvalues. The model is then solved for each of the six VSI device conduction modes using nodal equations and matching boundary conditions.

Typical results for startup from rest are shown in Figure 2. The transition from asynchronous to synchronous PWM at

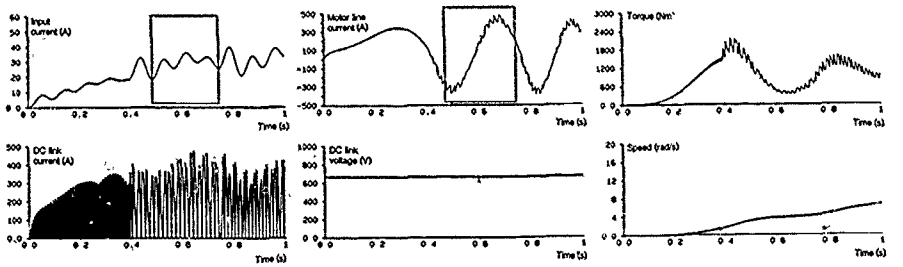


Fig. 2. Simulator outputs for one-second traction drive acceleration period.

about 0 4 s is clearly shown, as also are variations in input current and torque as the drive accelerates.

III. ON-LINE FFT GENERATION

The simulator has been implemented on a 68020-based parallel computer [2]. The task structure is shown in Figure 3. The procedure to initiate a FFT calculation is to initialise the system by keyboard command, and to define the time window of interest by selecting the starting instant, the sampling interval and the number of points. The master program then informs the calculator task that the FFT calculation is underway. The parallel structure of the simulator ensures the FFT and model calculations proceed simultaneously, although the FFT itself is calculated serially.

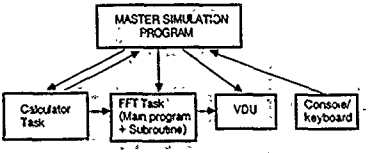


Fig. 3. Simulator task organisation.

Computation of the FFT is based on the Cooley-Tukey algorithm [3] developed by Brigham [4]. The algorithm implementation is achieved by a main program, which is a routine of the simulation model, with a subroutine to perform the Fourier analysis. The main program passes the data set (signal amplitude, total window angle) and the number of points (in binary) to the subroutine, where the sine/cosine value table is set up, and the calculations and sums performed, the results being returned to the main routine.

The procedure requires the input function to be represented in 2N points (Figure 4). An N-point transform is used to compute the real and imaginary parts, $X_r(n)$ and $X_i(n)$, of the discrete Fourier transform.

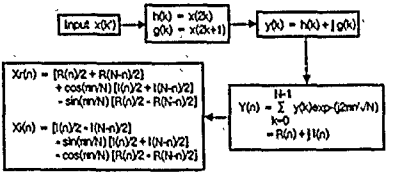


Fig. 4. FFT algorithm. x(k), h(k), g(k), R(n), I(n) are real; y(k), Y(n) are complex; k' = 0,1, . . . 2N-1, and k, n = 0,1, . . . N-1.

An advantage of the Cooley-Tukey algorithm is a reduction in the number of multiplications and additions in the computation. However this is at the expense of the need to de-scramble the output data. Thus after the values have been returned by the subroutine, the main routine separates the frequencies to return the completed harmonic analysis result to the simulation program for console display. The algorithm used to drive the results in natural order uses the decimation in time technique [4], where powers of exp-(j2π/N) are provided in the correct order needed for computation, thus eliminating the need for extensive storage tables.

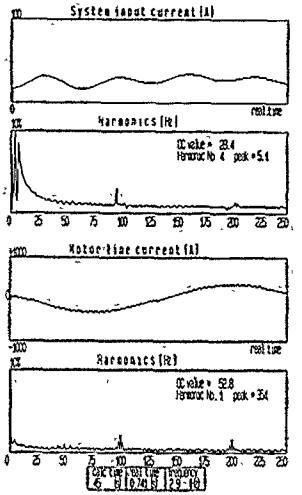


Fig. 5. System input and motor line currents and FFT spectra

The system input and motor line currents in Figure 2 have been windowed as shown and subjected to FFT analysis. The results, Figure 5, show the motor line current as almost sinusoidal, with more harmonics in the input current. Shown are the amplitude of the most significant harmonic, and the modulation frequency (2.9 Hz), real time (0.871 s), and calculation time (45 s).

IV. DISCUSSION AND CONCLUSIONS

Measurement of periodic signal parameters is generally subject to errors from spectral leakage and harmonic interference. Normally, interpolation algorithms are used to minimise these errors. Careful choice of sampling frequency and truncation envelope is thus required. The use of flat-top windows, as described by Salvatore and Trotta [5], is currently being investigated to reduce errors.

The simulator has proved to be a convenient, flexible tool which has enabled harmonic production at switching transitions between various forms of PWM to be investigated.

V. REFERENCES

[1] Hill R.J. and Huang F . On-line simulator for transient behaviour of inverter-fed induction motor traction drives. Proc. IMACS-TCI '90, Nancy, 19-21 Sept. 1990, pp. 417-22.
[2] Hill R.J. and Huang F.: Performance prediction of inverter-induction motor drives for DC-fed railways. 4th Int Conf. PEVSD, London, 17-19 July 1990, pp 529-33
[3] Cooley J.W. and Tukey J.W.. An algorithm for machine calculation of complex Fourier series. Math Computation, Apr. 1965, v. 19, pp. 297-301.
[4] Brigham E.O.. The Fast Fourier Transform Englewood Cliffs NJ: Prentice-Hall 1974.
[5] Salvatore L and Trotta A Flat-top windows for PWM waveform processing via DFT Proc IEE, Nov 1988, v 135B, n 6, pp. 346-61.

A new Method for direct digital Control of thyristors Converters. Presentation. Modelling.

Jean-Paul LOUIS, Ameziane DJERROUD, Claude BERGMANN
Laboratoire d'Electricité, Signaux et Robotique
(C.N.R.S. - U.A. D 1375).
Ecole Normale Supérieure de Cachan.
61 Avenue du Président Wilson. 94235 Cachan Cédex. France.

Abstract.

A new strategy with varying sequences for thyristors converters suitable for digital control is presented. Modelling (in sampled data sense) is given and a synthesis method is proposed.

Introduction: Analog versus Digital Control of Converter.

Analog control of thyristors converters (in rectifier or inverter mode) has a long history and is well known. Classical modelling and synthesis with linear modelling are often used, but rigourus modelling of these devices leads to non-linear, sampled data equations (/3/, /4/, /5/). The most remarquable property is that the converter realizes in fact an analog-digital conversion; the instant of sampling is identical with the instant of control, so the system has no delay. This is a good property with regard with the stability This is particulary important for fast regulations as current regulations which are here taken into consideration.

With digital control the problem becomes different. The designer must choose a strategy. Conceptually, the most easy is an "hybrid stategy" where the digital component "simulate" an analog control: with a frequency as high as possible, the digital component generates the same signals as the analog component. This strategy has some good properties (no delay, good stability), but the microprocessor is always computing the signals. So it as no time for other activities relating to high level and this is good only if the control frequency is effectively high with regard with the frequency of the converter (300 Hz in Europa). This strategy is not optimal but it has some advantages and is often used.

The strategy proposed here will employ some advantages of digital control: at each period of the converter the digital component makes the measurement one time and the control is computed also one time. So the digital component has time for other activitates But the questions are. Choice of the sampling instants, Determination of the contol algorithm, which is deduced from the properties of: The modelling of the system.

A SPECIFIC STRATEGY FOR THYRISTORS CONVERTER CONTROL.

For thysistors converters we must remark that at each instant it is possible to fire three different thyristors. With analog control the sequence is imposed and normally only one thyristors is accessible. For digital controllers the freedom is more important and the three thyristors are effectively eligible.

The most immediate strategy is the following: we choose the "natural firing instant" (cf./1/). The figure 1 shows that the delay t_r between the sampling instant and the effective control instant (the firing instant) depends on the firing angle:

$$t_r = t'_n - t_n = \Psi/\omega \qquad (1)$$

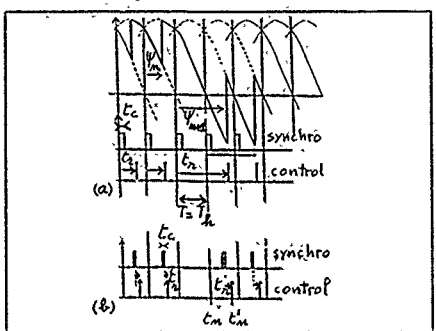


Fig.1:Direct digital control of a thyristor converter. (a) strategy with a delay which can be greater than the period. (b) strategy with a delay smaller than the period.

We see that this delay varies between 0 and three periods of the converter. The delay does not exist in analog control. And it is well known that a delay has bad effects on the stability of closed loop systems. This is particulary dangerous here because the delay can be greather than one period (see fig.1a). And when the delay is very large, the digital control can lead to regulation having a very poor stability margin.

That is the reason why we propose a new direct digital control with varying sequence which is choosen so that the delay is always smaller than a period of the controller. Fig.1b gives the principal of the control: the synchronization instant is "adaptative" and depends on the value of the firing angle (in the classical sense). Fig.2 defines the varying sequences We observe 4 different sequences.

(1) $0 \le \Psi < \pi/6$ (2) $\pi/6 \le \Psi < \pi/2$
(3) $\pi/2 \le \Psi < 5\pi/6$ (4) $5\pi/6 \le \Psi < \pi$

In this case, the delay is always smaller as one period, thus with simple algirithms (P.I. for example), we can have good dynamical behaviour (deadbeat respons, for example).

SAMPLED DATA MODELLING.

Then we propose a mathematical method to model the system in the sens of the sampled data system theory, and we deduce from this modelling a method for compute the parameters of the controller (by choosing the location of the poles). We consider small perturbation around the steady-state, and we note:

$$\Psi_n = \Psi_\infty + \delta\Psi_n$$
$$i(t_n) = i(n) = i_\infty + \delta i(n)$$

Thus a structural diagram of the system is given by fig.3a. the suppiy (thyristors converters and pulses generator) of a Resistance-Inductance (R-L) is controlled by a digital control. An equivalent scheme is given by fig.3b where we distinguish: the electrical time constant $T_e = L / R$ and the delay of the control:

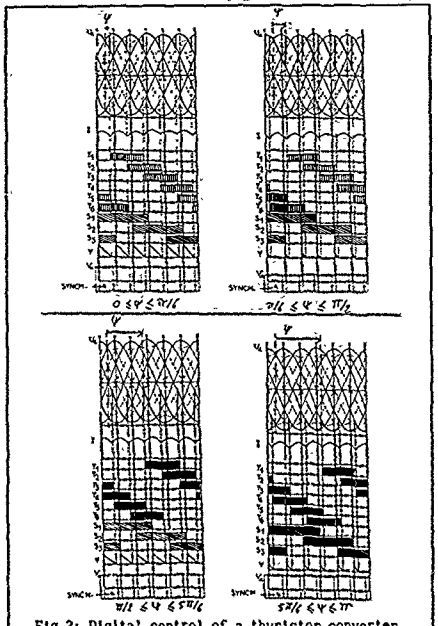


Fig.2: Digital control of a thyristor converter with modification of the firing sequence.

$t_r = t'_n - t_n$: algebraic calculus on rigourus (non-linear) equations gives the value of the converter gain (in sampled data sense)

$$K_e = -h(\psi_\infty) \qquad (2)$$

where $h(\psi_\infty) = 2.\sin \pi/6.\sin \psi_\infty$ is the discontinuity of the supply voltage at the firing instant; the term a_0 denotes the conversion coefficient between the output signal u_a of the controller and the firing delay.

$$a_0 = dt'_n / dv_a \qquad (3)$$

In practice, we will consider two output conversions.
- linear : $u_a = u_{am}\ (1 - 2\ \psi/\pi)$, $\quad a_0 = -\pi / 2.\omega.u_{am}$
- cosine : $u_a = u_{am} \cos \psi$, $\quad a_0(\psi_\infty) = -1 / u_{am}.\sin \psi_\infty$

The transfert function relative to small perturbations is:

$$\delta I(n+1) = e_0\ \delta I(n) + (\ e_0 / e_1).(-h / L).\delta t'_n \qquad (4)$$

with $e_0 = \exp(-T/T_e)$, $\quad e_1 = \exp(-t_r / T_e)$.

CONTROL OF THE MEAN CURRENT.

A problem is dus to the fact that on one period the current is measured only one time, so we have no sufficient information on the value of the current if it has a big ripple (this is specially important in discontinuous mode). Thus we propose to use the integral value (in analog sense) of the current with a reset to zero at the end of each period (cf./3/). Thus we can control the mean value of the current, which is in practice the value of interest.

Fig 4 present the structural diagram and the equivalent scheme of the current control with measurement of the mean value (k_{d1} and k_{d2} are the gains of the A/D converter. Fig.5 gives the shape of the current and of J (the current integral). The modelling of this system is the following. The current equation is the same as previous. The integral is directly defined in sampled data sense:

$$J(n+1) = \int_{t_n}^{t_{n+1}} i(t).dt \qquad (5)$$

It can be demonstrated that the equation relative to small perturbation around an operating point is

$$\delta J(n+1) = \alpha.\delta I(n+1) + \beta.\delta I(n) \qquad (6)$$

with $\alpha = T_e.(e_1/e_0 - 1)$, $\quad \beta = T_e.(1-e_1)$. Then the complete modelling of the system is given by (4) and by:

$$\delta u_a(n+1) = \lambda.\delta I(n) + \mu.\delta u_a(n) + K_p.K_1.\delta V_{ref}(n+1) \qquad (7)$$

with: $\lambda = -k_p.(e_0.(1+\alpha.K'_1) - (1-\beta.K'_1))$,

$\mu = 1 - \gamma(\psi_\infty).k_p.(1 + \alpha.K'_1)$

$\gamma(\psi_\infty) = (e_0/e_1).(-h(\psi_\infty)/L).a_0(\psi_\infty)$.

$k_p = K_p.k_{d1}.k_1$, $\quad K'_1 = K_1.k_{d2}/k_{d1}$.

The closed-loop transfert function is.

$$\frac{\delta I(z)}{\delta V_{ref}(z)} = \frac{\gamma(\psi_\infty).K_p.K_1}{z^2 - (e_0 + \mu).z + e_0.\mu - \gamma(\psi_\infty).\lambda} \qquad (8)$$

SYNTHESIS.

We can choose two roots equal to zero to obtain the deadbeat response. Thus the parameters must verify: $e_0 + \mu = 0$ and $e_0.\mu - \gamma.\lambda = 0$. And the solution is.

$$K'_1 = 1/\ T_e.(1-e_1.e_2); \qquad (9)$$

$$\gamma(\psi_\infty).k_p = (e_0/e_1).(1-e_1.e_0)/(1-e_0) \qquad (10)$$

We observe that K'_1 (and K_1) depends on the operating point (see fig.6) and K_p depends strongly of the operating point if the conversion factor is a constant (linear firing), but varies lightly if the conversion factor is a sine (cosine firing, see fig.6).

Fig.7 gives dynamical experimental behaviour of the system. The observed performances are good.

CONCLUSION.

We have proposed a solution to some problems related to digital control of thyristors converters, as: choice of the synchronization instants with adaptative sequences, and how to regulate currents with high ripple. We have proposed a precise sampled-data modelling and given formulas to optimize the dynamic behaviour. The modelling can be generalized to discontinuous mode. Implementation validates the theoretical previsions.

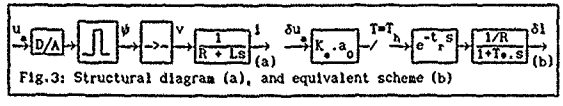
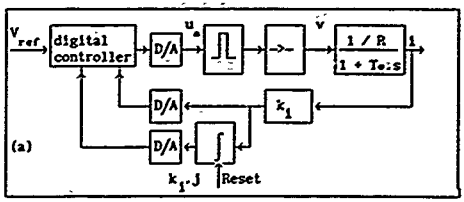


Fig.3: Structural diagram (a), and equivalent scheme (b)



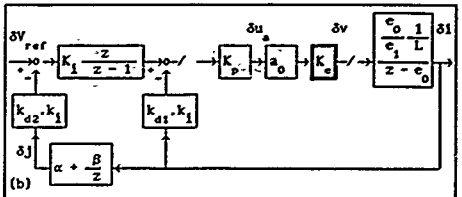


Fig.4 :schemes of a current control including a measurement of the mean value of the current.

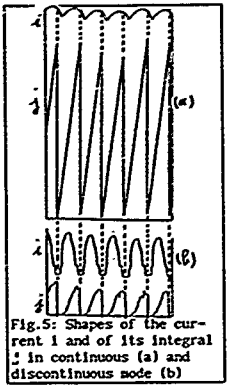


Fig.5: Shapes of the current i and of its integral $\frac{1}{\tau}$ in continuous (a) and discontinuous mode (b)

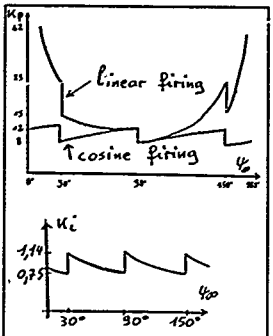


Fig.6 : Variation of K_i and K_p when the operating point varies. (mean value control).

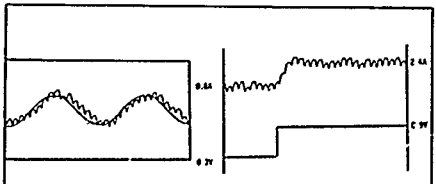


Fig. 7. Dynamic behaviours of the current regulation (mean value)

REFERENCES

/1/ K.P. GOKHALE, G.N. REVANKAR:" Microprocessor-controlled separately excited DC-motor drive system.", IEE Proc., vol. 129, Pt.B, N°6, nov. 1982, pp. 344-352.

/2/ T. OHMAE, T. MATSUDA, R. MASAKI, K. SAITO: "A Microprocessor-based current controller with an internal current-rate loop for motor drive". IEE trans. Ind. Appl.,vol IA-22, N°5, Sept./Oct. 1986, pp.805-811.

/3/ J.-P. FAVRE:"Microprocessor-based speed control of a DC-motor", Symp. Darastadt, Microelectronics in Power Electronics and Elec. Drive, 12-14 oct. 1982, 257-263.

/4/ J.-P. LOUIS: "Non-linear and linearized models for control systems including static converters". 3rd Symposium IFAC, "Control in Power Electr. and Elec. Drive",Lausanne, Sept. 1983, R.Zwicky, Pergamon-Press, pp. 9-16.

/5/ J.-P. LOUIS:"Application of a sampled data Modelling of static converters to the analysis and synthesis of certain regulations". 1rst IMACS-TC1 Symp., Liège (Belgium), 1984, "Electrical Machines and Converters, Modelling and Simulation", H.Buyse, J. Robert Ed., Elsevier Sc. Pub., North-Holland , pp. 139-146.

/6/ D.LEFEVRE, C.BERGMANN, J.P.LOUIS: Comparaison d'Algorithmes pour la Commande numérique directe de servomoteurs. Cas des Boucles de Courants. 5ème Colloque Moteurs Pas à pas, Nancy, 22-24 Juin 1988.

/7/ A. DJERROUD:"Commande numérique directe d'un convertisseur à thyristors en pont triphasé." Thèse de doctorat, Université de Paris VI, 22 juin 1988.

Modelling and Voltage Feedback Control of the Injected Current Predictive Modulator of a DC-AC Parallel Resonant Converter

J. Fernando Silva, B. V. Borges and J. Santana

I.S.T. / I.N.I.C., D E E C, Máquinas Eléctricas e Electrónica de Potência
Av. Rovisco Pais 1095 Lisboa Codex
PORTUGAL

ABSTRACT — No adequate model is known, to the authors, for the dynamic behaviour study of the injected current predictive modulator, for a resonant DC-AC converter. An approach, based into predictive control principles, is used to obtain a convenient open loop linear model of the DC-AC converter, with a parallel resonant high frequency link, suited to high power levels. Theoretical and experimental results concerning the synthesis and the performance of the voltage feedback regulator, using the proposed model, are presented.

1. INTRODUCTION

The use of the resonant principle in DC- AC power conversion has been increased in the last few years, leading to new, more sophisticated and efficient topologies. However, in most of them, the control process for the converter modulator is difficult to establish, due to the absence of a convenient open loop small signal model, suitable for the synthesis of the closed loop regulators.

One example is the injected current predictive modulator, employed to control the output current of a cycloconverter driven by a current fed parallel resonant bridge inverter. The predictive modulator controls the converter output current, but, in most environments, output voltage control is also needed. In order to achieve this, the modulator still requires additional output voltage regulators, which must be designed with the aid of a suitable small signal transfer function. The definition of this modulator transfer function presents the interesting problem of building a predictive control model, adequate for the output voltage regulator implementation, using simple linear feedback theory. The feasibility of this control method has already been demonstrated with experimental results presented by one of the authors [1].

2. INJECTED CURRENT PREDITIVE MODULATOR: PRINCIPLES OF OPERATION

The operation of this modulator can be briefly reviewed as follows. The whole DC-AC converter, including the output cycloconverter is a line commutated circuit. Hence semiconductor commutation is performed only at the oscillating voltage zero crossing points of the current fed parallel resonant bridge inverter, in order to reduce switching losses. Therefore, the output voltage waveform is made with half sinusoids of both polarities or null voltage (fig.1). Thus, during each half cycle of high frequency input voltage, the cycloconverter has only three possible output voltages or modes of operation: (1) full positive half cycle voltage ; (2) full negative half cycle voltage; (3) null voltage (freewheeling). In the beginning of each half cycle, one of this three operation modes is selected in order to inject, in the load and output filters, a certain desired current.

This is accomplished by the modulator, whose current predictor, at the end of each output voltage half period, calculates the predictable variation, for the next half cycle, of the cycloconverter output current for all the three operation modes. Current prediction is accomplished by calculating the three output current first time derivatives, due to the three possible output voltages. The operation mode for the next half cycle is defined by the comparison of the three predicted output currents and the reference or desired output current value (I_{ref}). The control circuitry selects the mode that will lead to the nearest current value, and a logic circuitry triggers the semiconductor devices, with a certain discrete trigger angle (σ), suitable to establish the desired circuit topology. This process is repeated each half cycle and the decision taken can not be changed until the end of the next half cycle.

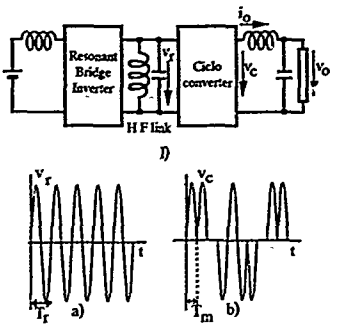


Fig.1 I) DC-AC converter with output cycloconverter and parallel high frequency link. II) a) Cycloconverter input oscillating voltage and b) output modulated voltage.

3. MODELLING THE INJECTED CURRENT PREDITIVE MODULATOR

The control circuit described is a good example of a discrete system, but none of the well known models for discrete systems is easy to use, due mainly to the difficult calculation of the steady state gain. The difficulty arises from the system complexity (inductive current source, parallel resonant LC, load and output filter, with a probable fourth order transfer function with a non predictive non current mode modulator), coupled to the current mode predictive modulator, and to the large fluctuations of the oscillating voltage amplitude of the parallel resonant circuit These large variations also prevent the use of a control method based on the principle of a programmed waveform PWM modulator [2]

However, being a discrete system, with the best possible transient performance that can ever be expected, the output current (i_o), will follow the reference current (I_{ref}) in the next (i+1) half cycle [3]. Therefore, it can be written:

$$(i_o)_{\sigma_{n+1}} = I_{ref}, \qquad (1)$$

where $(i_o)_{\sigma_{n+1}}$ is the output current due to trigger angle σ_{n+1}

From (1), $(i_o)_{\sigma_{n+1}} - I_{ref} = 0 = (i_\Delta)_{\sigma_{n+1}}$. Using the Taylor series expansion around the true solution σ_{n+1}, it is obtained [4]:

$$(i_\Delta)_{\sigma_{n+1}} = (i_\Delta)_{\sigma_n} + \frac{\partial (i_\Delta)_{\sigma_n}}{\partial \sigma_n}(\sigma_{n+1}-\sigma_n) + \frac{1}{2!}\frac{\partial^2 (i_\Delta)_{\sigma_n}}{\partial \sigma_n^2}(\sigma_{n+1}-\sigma_n)^2 + \ldots \qquad (2)$$

Taking only the first two terms of the series, approximating derivatives by finite differences and with some algebraic manipulation, equation (2) gives:

$$I_{ref} = (i_o)_{\sigma_n} + \frac{d(i_o)_{\sigma_n}}{dt} T_m, \qquad (3)$$

where $(i_o)_{\sigma_n}$ represents the actual output current due to trigger angle σ in half cycle n, and T_m is this half cycle time value. This

equation justifies why the predictive modulator, previously built, takes samples of $(i_o)_{\sigma_n}$, and of the input and output voltages. These samples allow the calculation, of the 3 possible first time derivatives of the output current and corresponding time increments to forecast the current $(i_o)_{\sigma_{n+1}}$.

Making $I_{ref} = g_m v_\varepsilon$, where v_ε is a reference voltage and g_m is a practical constant, calculated by the ratio $i_{o_{max}}/v_{\varepsilon_{max}}$, it is obtained:

$$(i_o)_{\sigma_n} = g_m v_\varepsilon - \frac{d(i_o)_{\sigma_n}}{dt} T_m . \qquad (4)$$

This equation shows that in the steady state the whole DC-AC converter is a transconductance amplifier ($i_o = g_m v_\varepsilon$). This conclusion can be supported by simulation and experimental results, shown in figure 2.

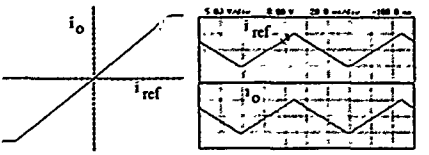


Fig.2 Steady state open loop transfer function:; a) Computer simulation b) Experimental result.

The small signal model can be obtained applying Laplace transform to the small signal equation from (3) which gives:

$$\frac{I_o(s)}{V_\varepsilon(s)} = \frac{g_m}{1+sT_m} . \qquad (5)$$

The previous equation means that the converter can be seen as a transconductance operational amplifier, with a dominant pole whose frequency is $1/(2\pi T_m)$.

4. SYNTHESIZING THE VOLTAGE FEEDBACK REGULATOR

The simple model obtained allows the use of conventional linear feedback control theory to implement the voltage regulator The block diagram of the circuit uses a P. I. regulator (fig 3), which allows zero steady state error and fast transient response.

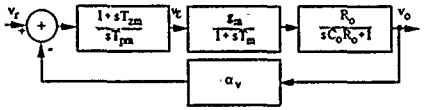


Fig.3 Block diagram of the voltage controlled system.

Cancelling the pole due to the load with the regulator zero, and assuming a damping factor $\zeta=\sqrt{2}/2$ it can be easily shown, that the P.I. regulator parameters are:

$$T_{zm} = C_o R_o \; ; \; T_{pm} = 2\,\alpha_v\, g_m R_o T_m \qquad (6)$$

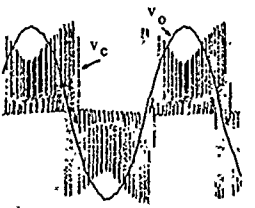


Fig.4 Experimental results: Output filter voltage v_o and cycloconverter output modulated voltage v_c.

Experimental results shown in figures 4 and 5 confirm the correct choice of the P.I. regulator altogether with its parameters, showing that the now proposed converter model, although very simple, is also very accurate. However, a discrete model should be derived from this simple linear continuous model, for determining the stability of the converter. This subject will be discussed in detail in a future work.

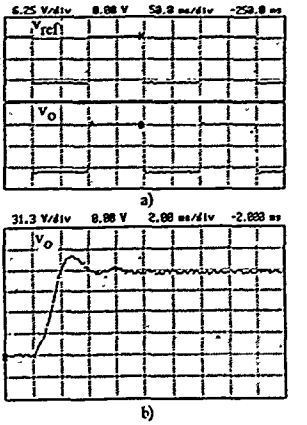


Fig.5 Experimental results: Reference voltage and output voltage a) Square wave input b) Rise time and Overshoot.

5. CONCLUSION

A suitable model was presented for the injected current predictive modulator, employed to control the output current of a cycloconverter driven by a current fed parallel resonant bridge inverter. The whole DC-AC converter is theoretically treated as a transconductance amplifier, and experimental results confirm the validity of this model and its usefulness to synthesize voltage feedback regulator with very good transient response.

REFERENCES:

[1]. Borges, B. V., Prediction Current Injected Control Applied To DC-AC Parallel Resonant Control, Proc. of the 15th IEEE Industry Electronics Conference, Phil. USA, 1989

[2]. Borges, B. V , Conversor Corrente Contínua Corrente Alternada com Andar Ressonante de Alta Frequência, Doctor Degree Thesis in Electrical and Computer Engineering, Instituto Superior Técnico, Lisboa, 1990.

[3]. Luo, F. L.; R. J. Hill;, Fast response and optimum regulation in digitally controlled thyristor converters, IEEE Trans on IA, vol. IA 22 nº 1, pp. 10-17, 1986.

[4]. Fernando Silva, J.; Controlo preditivo por simulação para rectificadores com eliminação de falhas de comutação, Doctor Degree Thesis in Electrical and Computer Engineering, Instituto Superior Técnico, Lisboa, 1989.

MODELLING AND CONTROL OF A PULSE WIDTH MODULATED HIGH POWER OPERATIONAL AMPLIFIER

FERNANDO SILVA, J.

I.S.T. / I.N.I.C. - D.E.E.C. Máquinas Eléctricas e Electrónica de Potência
Av. Rovisco Pais 1096 Lisboa Codex
PORTUGAL

ABSTRACT— A linear small signal model for the association of a half-bridge inverter and a current mode pulse width modulator, is introduced. The model is suitable to easily synthesize the feedback loops needed to obtain a four quadrant inverter, with GTO thyristors switched at 1KHz, which then performs as an operational amplifier. However, this continuous model is not convenient for the stability prediction, so it is also presented a discrete model, with the same steady state gain, and used in determining the converter stability. Theoretical and experimental results concerning both operational amplifier performance and stability, are presented.

1. INTRODUCTION

Voltage inverters (DC-AC power converters) are the main electronic sub-system of the now widely used uninterruptible power supply, and are also the fundamental system of most variable speed drives with induction motors. Although in most voltage inverters the control process is well established giving acceptable capabilities, recently there has been a demand for better reliability and performance. This fact led to the implementation of the control method, named PWM dual current mode instantaneous voltage feedback control [1]. Whilst a substantial amount of work has been invested in studying the intrinsic stability of the modulator principle, as far as the author is aware, little effort has been done both on the modelling of the modulator when included in a feedback controlled system and on the process stability study. This work tries to fill this gap, so adequate models for dynamic behaviour design and stability analysis of the current mode modulator controlled system are proposed.

A dominant pole continuous linear model, applied to the current mode modulator, is refered, due to its quasi-universal applicability to power converters, and to its simplicity, which allows a correct feedback control design, even for the non specialist. However, stability analysis theoretical results obtained with the continuous model do not match experimental ones, so the model is not suitable for stability studies. Therefore, a discrete model is also presented, and found to be well suited for stability analysis. However, it is more cumbersome for feedback regulator design, although it reproduces the same parameter values.

2. CONTINUOUS LINEAR MODULATOR MODEL AND INSTANTANEOUS VOLTAGE FEEDBACK REGULATION.

The dual current mode modulator compares a signal proportional to the inverter output filter and load current, with two voltage levels v_t+V_a and v_t-V_a, where v_t is the error voltage value at the modulator input and V_a is a constant voltage that must be larger than a voltage proportional to the half maximum output current ripple. A dedicated logic circuitry associates the comparison results with a fixed frequency two phase clock, deciding, in each period, which power semiconductors must be turned on or off, synchronously with the two clock phases, and which must be turned off or on upon the comparison results, thus varying the duty-cycle. In this way, this fixed frequency modulator allows four quadrant converter operation and provides output current auto-clipping capability.[1].

The association, power inverter-dual current mode PWM modulator with instantaneous output current value sampling, can be approximately described in the steady state, continuous operation, by the equation, $I_o = g_m v_t$, where I_o is the output current mean value per period, v_t is the control voltage, and g_m is the association related constant (transconductance). Thus, the converter is equivalent to a voltage (v_t) controlled current source (or sink) with transconductance g_m or an operational transconductance amplifier (OTA).

The open loop dynamic model can be obtained assuming the statistical mean value for the modulator discrete delay, T_m ($T_m=T_{PWM}/2$ where T_{PWM} is the period of the PWM carrier frequency). This gives $i_o = g_m v_t e^{-sT_m}$, as the dynamic model. Approximating the exponential by the first two terms of the e^x Taylor series expansion, which is valid for $sT_m<0.6$, it is obtained a dominant pole continuous linear model (industrial model) [2], for the association inverter-modulator:

$$i_o(s) = \frac{g_m}{1+sT_m} v_t(s) . \qquad (1)$$

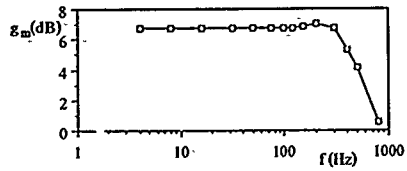


Fig 1 Experimental open loop frequency response. Note that the frequency of the theoretical pole is closed to the frequency of the experimental one (T_{PWM} = 1ms).

These assumptions are validated experimentally by figures 1 and 2

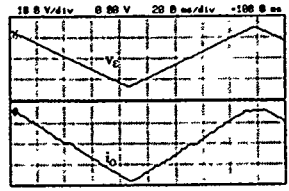


Fig.2 Open loop experimental results. Reference voltage v_t and output current i_o.

To control the inverter output voltage V_o, the use of a proportional integral (P. I.) regulator, handling the error between the voltage reference v_r and the output voltage instantaneous value v_o, guarantees zero steady state error and sufficiently fast transient response. The whole inverter block diagram is shown on figure 3

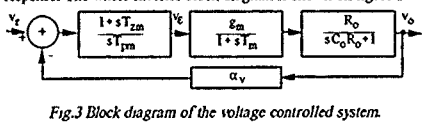


Fig.3 Block diagram of the voltage controlled system.

Cancelling the pole, dependent on the load resistance R_o and C_o of the LC output filter, with the zero of the P.I. regulator:

$$T_{zm} = C_oR_o , \qquad (2)$$

a 2nd order closed loop transfer function can be written

$$\frac{v_o(s)}{v_r(s)} = \frac{1/\alpha_v}{\left(\frac{T_{pm}T_m}{\alpha_v g_m R_o}\right)s^2 + \left(\frac{T_{pm}}{\alpha_v g_m R_o}\right)s + 1} , \qquad (3)$$

where the damping factor must have a value close to $\xi = \sqrt{2}/2$, to

obtain low overshoot and fast response. Thus:

$$T_{pm} = 2\alpha_v g_m R_o T_m \quad (4)$$

The equations (2) and (4), together with the proper tuning of the LC output filter for 40 dB ripple carrier rejection [3], allow the design of a power operational amplifier with the voltage gain set by $1/\alpha_v$ (20dB), maximum power bandwidth of 120 Hz (DC to 120 Hz), small signal bandwidth of 160 Hz (fig. 4) and good static and dynamic performances, which can be seen in the experimental results presented (fig. 5 and 6).

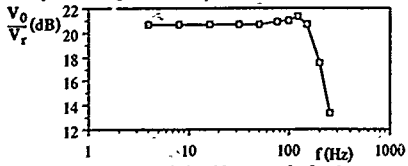


Fig.4 Experimental closed loop transfer function.

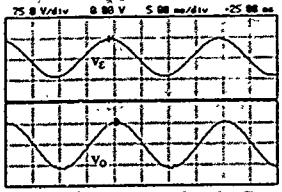


Fig.5 Closed loop experimental results: Sinusoidal reference v_ϵ and output v_o voltages.

The usefulness of the simple continuous model for feedback synthesis, justifies its presentation in this work (for divulgation purposes), as it can be a first order model for most power converters.

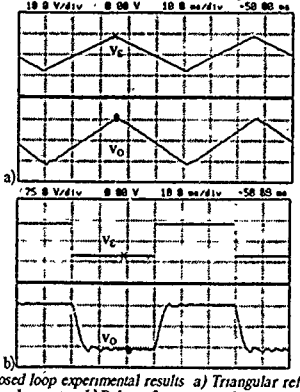


Fig 6 Closed loop experimental results a) Triangular reference v_ϵ and output v_o b) Pulse reference v_ϵ and output v_o response.

3. MODULATOR DISCRETE IMPULSE MODEL AND STABILITY ANALYSIS

To determine a linear incremental model which accurately accounts for the discrete action of the dual current modulator principle, let us consider only infinitesimal perturbations Δv_ϵ, that originate also infinitesimal perturbations in the duty-cycle and hence on the inverter output pulsed voltage Δv_s. These perturbations have the form of infinitesimal narrow pulses, so they can be replaced by δ functions (Dirac impulses) multiplied by the pulse area. The incremental output voltage of the power inverter can, thus, be modelized by an ideal sampler (fig 7), which samples Δv_ϵ at the commutation instant. The perturbations on the output current mean value, linked to these pulses, are delayed by the output LC filter inductor. Then, the ideal sampler must be followed by a continuous lag with gain g_m and time constant T_{PWM}. The sampling frequency is the fixed frequency of the two phase clock ($f_{PWM} = 1/T_{PWM}$), and it will be assumed constant. This is a valid assumption, if there is only interest in determining the conditions leading to the onset of instability, which is the normal situation.

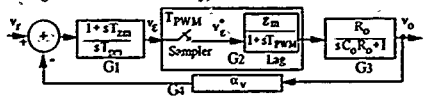


Fig.7 Discrete impulse model of the voltage controlled system.

A discrete Z transform based theoretical study shows that the discrete open loop transfer function $TF^*(z)$ is (fig.7):

$$TF^*(z) = G_1G_2G_3G_4^*(z) = K\frac{z(z-\beta)}{(z-1)(z-\alpha)(z-\beta)}, \quad (5)$$

where $K = g_m\alpha_v T_{zm}(1-\alpha)/T_{PWM}C_o$, $\beta = e^{(T_{PWM}/C_oR_o)}$ and $\alpha = e^{(-1)}$. Equating the expression (5) to minus one (using the modified Routh-Hurwitz criterion), the critical gain K_c can be calculated, and by root locus analysis technics, the instability frequency ω_c, can be determined: the root locus (fig 8a) crosses the unitary circle in the point (-1,0), then $\omega_c T_{PWM} = \pi$. Therefore:

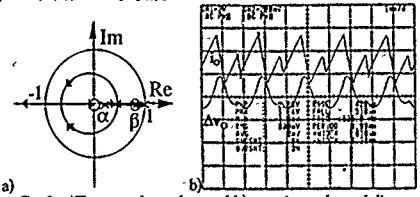


Fig 8 a)Theoretical root loci and b) experimental instability

$$K_c = 2(\alpha+1) \; ; \quad \omega_c = \frac{\omega_{PWM}}{2} \quad (6)$$

Instability at this frequency (half the sampling frequency) or Nyquist instability ,verified in practice (fig. 8b), is due to the system discrete nature, but is not predicted by the continuous linear model presented in section 2. This model predicts oscillations at a gain dependent frequency always bellow half of the sampling frequency.

4. CONCLUSION

This work presents two models for the dual current mode instantaneous voltage feedback inverter. For the design of the P.I. regulator, the whole DC-AC converter is theoretically best treated as a transconductance amplifier with a dominant pole. Experimental results confirm the validity of this model and its usefulness to synthesize voltage feedback regulator with very good transient response. Experimental results also show good performance levels due to the instantaneous feedback dual current mode PWM modulator and voltage feedback P.I. regulator. Nevertheless, this model predicts wrong critical gains and oscillating frequency.

For studying the closed loop system stability, the DC-AC converter is best treated as an ideal sampler followed by a lag. This model enable the correct calculation of the critical gain and the oscillation frequency, which is verified in practice, but requires a great mathematical effort to evaluate the P.I. regulator parameters.

REFERENCES:

[1]. Anunciada, A. V , M. M. Silva, A new current mode control process and applications, PESC Record, pp 1-12,1989.

[2]. Fernando Silva, J , Conversor electronico de potência com tiristores GTO para emulação de geradores de tensão e filtros, ENDIEL 89 ST1,pp. 227-238, 1989.

[3]. Fernando Silva, J., Controlo preditivo por simulação para rectificadores com eliminação de falhas de comutação, Doctor Degree Thesis in Electrical and Computer Engineering, Instituto Superior Técnico, Lisboa, 1989.

DIGITAL SIMULATION OF THE PWM VOLTAGE CONVERTER CONNECTED TO THE AC MAINS

P. VERDELHO AND G. D. MARQUES
IST/INIC, Secção de Máquinas Eléctricas e de Electrónica de Potência,
Av. Rovisco Pais 1096 Lisboa Codex, Portugal

Abstract - This paper presents detailed and global models for the simulation of the PWM voltage converter connected to the AC Mains. A personal computer program is introduced, allowing detailed and global simulation with system or with Park's coordinates. The numerical integration method used is an explicit Euler-Backward method. Theoretical results obtained with the detailed and global models and experimental results are presented and compared.

I. INTRODUCTION

Conventional thyristor AC/DC converters are limited in performance due to the line-commutation process. The advent of high power, high frequency semiconductor switches, with gate controlled fast turn-off capability (GTO's), allows the use of forced commutation and therefore PWM techniques can be applied to these converters [2].
PWM converters can be classified in two general types: current converters and voltage converters [4]. Voltage converters have useful characteristics [3] that can be summarized as:

- near sinusoidal current waveforms
- AC four quadrant operation
- DC link unidirectional voltage
- bidirectional power transfer capability by reversing the flow direction of the DC link current.

This paper presents some models for the simulation of the PWM voltage converter connected to the AC mains, fig1.

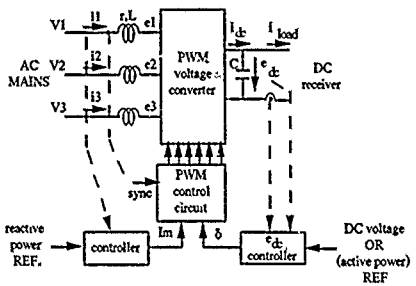


Fig 1. Voltage converter connected to the AC mains

Several applications are possible, specially in AC/DC conversion, where the DC receiver needs a supply with DC voltage characteristics. With the new high power, gate turn-off switches, applications on High Voltage Direct Current (HVDC) transmission [3] will be feasible.
The PWM control circuit, described in [4], is composed of analog and digital circuits. The inputs of this system are: 1) the modulation index I_m, and 2) the power angle δ measured between the mains and the AC fundamental converter voltages. Two detailed and two global models are presented. The first detailed model, in system state variables, is quite classical. A second detailed model, in Park's coordinates, is obtained from the first model by a Park's variables transformation. Global models (in state and in Park's coordinates) are obtained neglecting the harmonic content.
The paper presents experimental results obtained from a laboratory prototype (1 KVA). It is shown that experimental and theoretical results are in good agreement.

II. SYSTEM MODELING

A. System coordinates

The first detailed model, in system state variables, is a classical model. The PWM voltage converter is seen as a system with three transfer functions relating the e_{12},e_{23} and e_{31} line voltages with the DC voltage e_{dc}. The corresponding equations are:

$$e_{12}=(g_1-g_2)\,e_{dc} \;;\; e_{23}=(g_2-g_3)\,e_{dc} \;;\; e_{31}=(g_3-g_1)\,e_{dc} \qquad (1)$$

The transfer functions g_1, g_2 and g_3 are non-linear functions that can assume the values 0,1 according to the state of each converter arm, which are dependent of the PWM control circuit. Thus, they are time functions of the modulation index I_m, and of the power angle δ. Equation (1) can be written using phase voltages The transformed equation is:

$$e_1=f_1\,e_{dc} \;;\; e_2=f_2\,e_{dc} \;;\; e_3=f_3\,e_{dc} \qquad (2)$$

$$i_{dc}=f_1\,i_1+f_2\,i_2+f_3\,i_3 \qquad (3)$$

where:

$$f_1=\frac{2g_1-(g_2+g_3)}{3} \;;\; f_2=\frac{2g_2-(g_1+g_3)}{3} \;,\; f_3=\frac{2g_3-(g_1+g_2)}{3} \qquad (4)$$

The introduction of the three inductor equations of the AC side and the capacitor equation of the DC side gives the state model (5). In this model only the two independent AC currents are needed.

$$\begin{bmatrix} \frac{di_1}{dt} \\ \frac{di_2}{dt} \\ \frac{de_{dc}}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{r}{L} & 0 & -\frac{f_1}{L} \\ 0 & -\frac{r}{L} & -\frac{f_2}{L} \\ \frac{2f_1+f_2}{C} & \frac{2f_2+f_1}{C} & 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ e_{dc} \end{bmatrix} + \begin{bmatrix} \frac{V_1}{L} \\ \frac{V_2}{L} \\ -\frac{i_{load}}{C} \end{bmatrix} \qquad (5)$$

B. Park's coordinates

A second detailed model, in Park's coordinates, is obtained from (5) by a Park's transformation of variables (6):

$$\begin{bmatrix} \frac{di_d}{dt} \\ \frac{di_q}{dt} \\ \frac{de_{dc}}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{r}{L} & \omega & -\frac{f_d}{L} \\ -\omega & -\frac{r}{L} & -\frac{f_q}{L} \\ \frac{f_d}{C} & \frac{f_q}{C} & 0 \end{bmatrix} \begin{bmatrix} i_d \\ i_q \\ e_{dc} \end{bmatrix} + \begin{bmatrix} \frac{V_d}{L} \\ \frac{V_q}{L} \\ -\frac{i_{load}}{C} \end{bmatrix} \qquad (6)$$

f_d and f_q are computed from f_1, f_2 and f_3 in the same way as V_d and V_q are computed from V_1, V_2 and V_3

C. Global models

The two global models (in state and in Park's coordinates) are obtained from the corresponding detailed models by neglecting the harmonic content. In the state variable model, in steady state, f_1 and f_2 are sinusoidal functions of time, whose amplitude is proportional to the I_m value and whose phase is a function of the power angle δ
In the Park's model f_d and f_q are functions of I_m and δ

III. THE SIMULATION PROGRAM

The program implemented allows global and detailed simulation in systems or in Park's coordinates. To integrate equations 5 and 6; two explicit Euler-Backward formulas (7 and 8) are used:

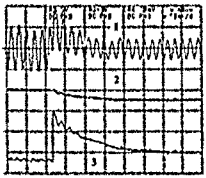
$$\begin{bmatrix} i_1 \\ i_2 \\ e_{dc} \end{bmatrix}_{t+dt} = \begin{bmatrix} 1+\frac{r}{L}dt & 0 & \frac{f_1}{L}dt \\ 0 & 1+\frac{r}{L}dt & \frac{f_2}{L}dt \\ -\frac{2f_1+f_2}{C}dt & -\frac{2f_2+f_1}{C}dt & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} i_1 \\ i_2 \\ e_{dc} \end{bmatrix}_t + \begin{bmatrix} \frac{V_1}{L}dt \\ \frac{V_2}{L}dt \\ -\frac{I_{load}}{C}dt \end{bmatrix}_{t+dt} \right) \quad (7)$$

$$\begin{bmatrix} i_d \\ i_q \\ e_{dc} \end{bmatrix}_{t+dt} = \begin{bmatrix} 1+\frac{r}{L}dt & -\omega\,dt & -\frac{f_d}{L}dt \\ \omega\,dt & 1+\frac{r}{L}dt & \frac{f_q}{L}dt \\ -\frac{f_d}{C}dt & -\frac{f_q}{C}dt & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} i_d \\ i_q \\ e_{dc} \end{bmatrix}_t + \begin{bmatrix} \frac{V_d}{L}dt \\ \frac{V_q}{L}dt \\ -\frac{I_{load}}{C}dt \end{bmatrix}_{t+dt} \right) \quad (8)$$

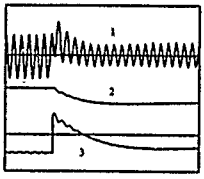
The system is simulated with a fixed modulation index $\hat{m}$ (e.g. without the reactive power controller). The e_{dc} controller is a PI designed to allow low overshoot and fast response. Equations corresponding to this controller are integrated by the Euler Forward method.

IV. RESULTS

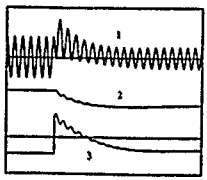
The results of the Park's model and the system state model are equivalent. In both cases, the global models are five times faster than detailed models. The numeric stability is very good.
The DC voltage controller parameters were selected in order to have a natural frequency ω_n=10 rad/s and a damping ratio ξ=0.7, in closed loop. For DC voltage control, experimental and theoretical results obtained with both models are shown in fig 2 and 3. If the controller parameters were modified in such a way that the natural frequency ω_n is increased, a malfunction occurs causing DC components in AC line currents, fig4. The global model also confirms this malfunction.



(a) Experimental



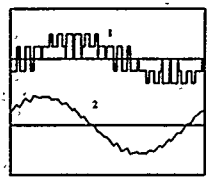
(b) Detailed Simulation



(c) Global Simulation

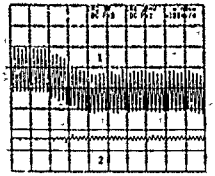
Figure 2. Response to a DC voltage reference variation (150→ 120V)
1 - AC line current 2 - DC voltage 3 - Power angle

(a) Experimental

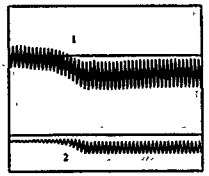


(b) Detailed Simulation

Figure 3. Steady state AC variables
(converter phase voltage, line current)



(a) Experimental



(b) Detailed Simulation

Figure 4. Response with improper controller parameters.
(line current, power angle)

V. CONCLUSION

Personal computer algorithms for the simulation of the voltage converter connected to the AC mains were presented. Two detailed and two global models have been presented, whose equations were integrated by an explicit backward method. The results of simulation were compared with experimental results obtained from a laboratory circuit test.
The models and the program presented are well adapted to the experimental prototype, being a valuable tool for system dynamics and regulation studies.

VI. REFERENCES

[1] Verdelho, P. ; (1990), "Ondulador de tensão trifásico ligado à rede de energia para o controlo de potência activa e reactiva", MSc. Thesis, IST, Lisbon

[2] Ohnishi, T.; Okitsu, H.; (1983), "A novel PWM technique for three phase inverter/converter", IPEC, vol. 1, pp.384-395.

[3] Ooi, Boon Teck; Wang, Xiao; (1990), "Voltage angle lock loop control of the boost type PWM converter for HVDC application", IEEE trans. Power Electronics, vol. 5, n.2, April, pp.229-235.

[4] Verdelho, P.; Marques, G.; (1990), "General control system of the PWM current/voltage converter connected to the AC mains", accepted to MELECON'91, Ljubljana.

THE GENESIS OF UNCHARACTERISTIC HARMONICS ARISING FROM SIX-PULSE AC/DC POWER CONVERTERS

L. Pierrat*,**, Y. J. Wang*, R. Feuillet*

* Laboratoire d'Electrotechnique de Grenoble
CNRS UA-355, BP. 46, Domaine Universitaire
38402 St. Martin d'Hères, France

** Electricité de France
Division Tech. Générale, 37 Rue Diderot
38040 Grenoble, France

Abstract - Uncharacteristic harmonic currents are generated by a six-pulse power converter due to imperfections in supply or control system. A quantitative analysis of the effects of supply system imperfections (i.e.; voltage unbalance and asymmetrical commutation reactances) upon the production of uncharacteristic harmonics is presented. Two unbalance factors (UBF_V and UBF_X) are proposed to characterize the imperfections so that the mathematical relation between harmonic intensity and the degree of unbalances can be established and reviewed.

I. INTRODUCTION

Although uncharacteristic harmonic currents produced by static converters are often neglected because of their relatively insignificant quantities compared with that of characteristic harmonics, the impedances at uncharacteristic harmonic frequencies may be quite high since filters are seldom provided for uncharacteristic harmonics. Therefore, significant uncharacteristic voltage harmonics may appear at the point of common coupling, especially when supply system possesses some degree of unbalance in voltage or impedances. Studies dealing with uncharacteristic harmonics of six-pulse converters apply complicated formulations to calculate the converter current during commutation intervals. This necessitates intensive computational effort for harmonic analysis [1-3]. Direct simulation of converter operation [4], although efficient and accurate as a tool for transient analysis, is not efficient to analyse current spectra in steady state. On the other hand, available analytical harmonic models [5] do not take into account the uncharacteristic harmonics. In this paper, a general analytical model with simplified commutation current for the determination of both characteristic and uncharacteristic harmonic currents of a six-pulse converter under unbalanced voltage and commutation reactances, is developed. Formulation for unbalance factors of voltage and reactance is also given. The formulation permits to establish the mathematic relationship between the unbalances (of voltage and reactances) and resulting uncharacteristic harmonics. Finally, the effects of the two kinds of unbalances are compared.

II. HARMONIC MODELLING [6]

In Fig. 1, the trapezoidal wave indicates the current waveform and the square wave represents its derivative. Note that current during commutation is approximated by a straight line. The commutation angles μ_1 and μ_2 differ from each other due to unbalanced commutation reactances. The current harmonics can be calculated as far as current waveform can be defined. This implies to define μ_1, μ_2, θ_1, θ_2, θ_3 and θ_4. The calculation of μ_1 and μ_2 gives no difficulties if the effect of voltage unbalance on μ is negligible [5].

$$\mu = \cos^{-1}[\cos\alpha - (x_1+x_2)I_d/(V\sqrt{6})] - \alpha \qquad (1)$$

where α indicates firing angle, V phase voltage in rms value, I_d dc current, and x_1 and x_2 reactances concerning the corresponding commutation. Whereas, to find θ_1, θ_2, θ_3 and θ_4 needs careful considerations for the effect of voltage unbalance upon line-voltage zero crossings. Fig. 2 shows an efficient method of determing the angles by which the zero crossings of line voltage deviate from their normal values due to voltage unbalance. A triangle formed by U_{ab}, U_{bc}, U_{ca} and an equilateral triangle with U_{bc} as one of its sides are shown. Taking U_{ab} as reference, δ_{ab} and δ_{ca} represent the deviation angles for U_{ab} and U_{ca}, respectively. The law of cosines gives:

$$\delta_{ab} = \cos^{-1}[(U_{bc}^2+U_{ab}^2-U_{ca}^2)/(2U_{ab}U_{bc})] - \pi/3 \qquad (2.1)$$

$$\delta_{ca} = \pi/3 - \cos^{-1}[(U_{bc}^2+U_{ca}^2-U_{ab}^2)/(2U_{ca}U_{bc})] \qquad (2.2)$$

Hence, θ_1 and θ_2 can be written as:

$$\begin{bmatrix} \theta_{1i} \\ \theta_{2i} \end{bmatrix} = (\alpha+\beta_i)\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \delta_{1i} \\ 2\pi/3+\delta_{2i} \end{bmatrix} \qquad (3)$$

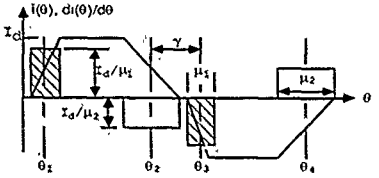


Fig. 1 AC current waveshape and its derivative

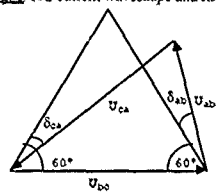


Fig. 2 Determination of deviation in zero crossings of line voltage

where i=a, b or c according to the phase specified. β indicates the phase shift between three phases. $(\beta_a, \beta_b, \beta_c) = (-\pi/3, \pi/3, \pi)$, $(\delta_{1a}, \delta_{1b}, \delta_{1c}) = (\delta_{ca}, \delta_{ab}, 0)$; $(\delta_{2a}, \delta_{2b}, \delta_{2c}) = (\delta_{ab}, 0, \delta_{ca})$. Because of the half period symmetry of current waveform, θ_3 and θ_4 are expressed as:

$$\theta_{3i} = \theta_{1i} + \pi \;;\quad \theta_{4i} = \theta_{2i} + \pi \qquad (4)$$

Note that I_d, μ_1, μ_2, θ_1, θ_2, θ_3 and θ_4 uniquely determine the current waveform. Consequently, the application of Fourier analysis to the waveform gives the harmonic components of line current. However, direct Fourier analysis on the trapezoidal current wave is a tedious work. A wiser method is to find the Fourier series of $di(\theta)/d\theta$ and then integrate the obtained series. Thus, the n^{th} harmonic current in can be written in a compact formula:

$$I_n = (I_d/\pi)[a_{n1}(e^{-jn\theta_1} - e^{-jn\theta_3}) - a_{n2}(e^{-jn\theta_2} - e^{-jn\theta_4})] \qquad (5)$$

where $a_{n1} = [\sin(n\mu_1/2)]/(n\mu_1/2)$; $a_{n2} = [\sin(n\mu_2/2)]/(n\mu_2/2)$; n = 1, 3, 5,..... For even n, $I_n = 0$. It can be shown that when the supply system is perfectly balanced, a_{n1} and a_{n2} are equal, $\theta_2 - \theta_1 = \theta_4 - \theta_3 = 2\pi/3$, and the harmonic currents of orders 3, 9, 15, 21, ... are null (i.e., the uncharacteristic harmonics do not exist). Note that good agreement between the current spectra calculated by Eq. (5) and by digital simulations has been obtained to validate the model.

III. FORMULATION OF UNBALANCES

Voltage unbalance and reactance unbalance are formulated in order

to relate uncharacteristic harmonics with unbalance factors UBF_v and UBF_x.

A. Voltage Unbalance - Unbalance factor of three-phase voltage UBF_v is defined as the ratio of negative sequence component to positive sequence component. Or:

$$UBF_v = V^-/V^+ = \tau_v e^{j\theta_v} \quad (6)$$

The magnitude τ_v and phase θ_v can be more easily calculated from the magnitudes of line voltages by solving the equations [7],

$$(X^2+Y^2-X-Y-XY+1)/(X+Y+1)^2 = (\tau_v{}^2)/4 \quad (7)$$

$$(\sqrt{3})(X-Y)/(X+Y-2) = \tan\theta_v \quad (8)$$

$$X = U_{ab}/U_{bc} \quad Y = U_{ca}/U_{bc} \quad (9)$$

$$\text{sign}(\sin\theta_v) = \text{sign}(X-Y) \quad (10)$$

With Eq (7)-(10), τ_v and θ_v have one-to-one correspondance with X and Y If U_{bc} is taken as reference (1 0 p u), U_{ab} and U_{ca} can be solved from τ_v and θ_v.

B. Commutation Reactance Unbalance - The unbalance factor of reactances is defined similar to UBF_v. That is, $UBF_x = \tau_x e^{j\theta_x}$, where τ_x and θ_x are determined by Eqs. (7)-(10) by substituing the three-phase reactances x_1, x_2 and x_3 for U_{ab}, U_{bc} and U_{ca}, and τ_x and θ_x for τ_v and θ_v. It is not difficult to show that (x_1, x_2, x_3) and $(\tau_x, \theta_x, x_{ave})$ have one-to-one correspondance where $x_{ave} = (x_1 + x_2 + x_3)/3$. That is, from τ_x, θ_x and x_{ave}, the commutation angles μ_1, μ_2 of line current can be found.

C. Relation between Harmonics and Unbalance Factors - The mathematical development of UBF_v and UBF_x allows to calculae I_n from τ_v, θ_v, τ_x and θ_x, provided that x_{ave} is given and U_{bc} is taken as reference. In fact, in Eq. (5), it can be seen that UBF_v affects mainly θ_1, θ_2, θ_3 and θ_4, and UBF_x affects a_{n1} and a_{n1}. The seperation of the effects of UBF_v and UBF_x is useful in the following assessment. Moreover, as the uncharacteristic harmonic currents are unbalanced, it is convenient to define a harmonic intensity indicator σ_n that reflects the effect of three-phase harmonic currents. σ_n is defined as:

$$\sigma_n = 100 \cdot \sqrt{[(I_{an}{}^2 + I_{bn}{}^2 + I_{cn}{}^2)/(I_{a1}{}^2 + I_{b1}{}^2 + I_{c1}{}^2)]} \quad (11)$$
$$n = 3, 9, 15, 21...$$

It is clear that the relation between σ_n and unbalance factors is described implicitly by Eqs. (1)-(11).

IV. EFFECTS OF UNBALANCES ON σ_n

Both unbalance factors are composed of their magnitudes (τ_v, τ_x) and phases (θ_v, θ_x). Although not included in this paper, it can be shown that σ_n is a strong function of τ_v and τ_x but a weak function of θ_v and θ_x. Furthermore, most industrial standards set limites for τ_v and τ_x but not for θ_v and θ_x. Therefore, it is reasonable to concentrate our study on the influences of τ_v and τ_x. Fig. 3 shows the variation of σ_3 versus τ_v, with τ_x as the parameter. Note that for n>3, σ_n is always bounded by σ_3. The values given in Fig. 3 are computed under the following conditions: V=1.0 pu; I_d=1.0 pu; α=30°; x_{ave}=0.1 pu; θ_v=60°; and θ_x=30°. Note that τ_x=40%, θ_x=30° and x_{ave}=0.1 pu correspond to x_1=0.11 pu, x_2= 0.1 pu and x_3= 0 06 pu, which is a unbalance unrealistically large Fig. 3 clearly illustrates that, when τ is very small, σ_3 is dominated by τ_x As τ_v increases to the value where the ratio τ_x/τ_v is less than 10, the effect of τ_x is replaced by τ_v. It shoud be noted that the curve of τ_x=0 is above other curves when τ_v is large. This is indeed the attenuation effect of commutation reactances.

V. CONCLUSIONS

An analytical model for the determination of both characteristic and uncharacteristic current of a six-pulse converter under voltage and reactance unbalance has been developed. The model resolves the causes of the production of uncharacteristic harmonics into functional (the deviation of θ_1, θ_2, θ_3 and θ_4 due to voltage unbalance) and structural (the asymmetry of μ_1 and μ_2 due to reactance unbalance) perturbations. An application of the model, the relation between two unbalance factors and a harmonic indicator σ_n, is presented. It is concluded that the model is useful for clear explanation of the genesis of uncharacteristic harmonics and is applicable to the harmonic problems concerning six-pulse converters.

VI. REFERENCES

[1] Arun G. Phadke & James H. Harlow, "Generation of Abnormal Harmonics in High-Voltage AC-DC Power Systems", IEEE Trans. on PAS, Vol. PAS-87, No. 3, Mar 1968, pp 873-883

[2] J. Reeve & P. C. S. Krishnayya, "Unusual Current Harmonics Arising from High Voltage DC Transimission", IEEE Trans. on PAS, Vol. PAS-87, No. 3, Mar 1968, pp.883-893

[3] J. Reeve, J A. Baron & P. C. S. Krishnayya, "A General Approach to Harmonic Current Generation by HVDC Converters", IEEE Trans. on PAS, Vol. PAS-88, No 7, July 1969, pp. 989-895

[4] R. H. Kitchin, "Computer Simulation of Harmonics Interaction Effects between AC/DC Converters", Proc. of Intern. Symposium on EECPS, CAPRI, May 24-26 1989, Cassino, Italy

[5] A David Graham & Emil T Schonholzer, "Line Harmonics of Converters with DC-Motor Loads", IEEE Trans. on IA, Vol. IA-19, No 1, Jan/Feb 1983, pp. 84 - 93

[6] L. Pierrat, Y. J. Wang & R. Feuillet, "Analytical Study of Uncharacteristic harmonics resulting from AC/DC Converters", IEEE ICHPS IV, Budapest, Oct. 1990

[7] L. Pierrat & J. P. Meyer, "Unbalance Factor, it is as simple as ABC", RGE (Revue Générale d'Electricité), N° 6, pp. 18 - 26, June 1987

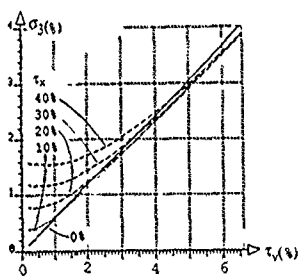


Fig. 3 σ_3 as functions of τ_v

COMPARISON BY SIMULATION BETWEEN ANALOGICAL AND OPTIMAL CONTROL OF STATIC VAR COMPENSATOR

M. AMOROUAYECHE
Sonelgaz ALGER
Algerie

C. COUNAN
Electricité de France
Dir. des Etudes et Recherches
1, Av. du Général de Gaulle
92141 CLAMART - France

J.M. KAUFFMANN
G.R.G.T.
IGE Parc Technologique
2, avenue Jean Moulin
90000 BELFORT - France

Abstract The aim of this paper is to show that a stabilizing signal proportional to the transmission line power flow variations added in the SVC voltage regulator, increases the dynamic stability of power systems.

Two kinds of signals are studied here; the design of the first one uses an analogical method and the second signal is achieved by using an optimal adaptive control. This last gives good results for small disturbances. In case of high magnitude perturbations, the analogical method can be more efficient.

Introduction

The integration of a far away power plant to an existing network creates always stability problems. To solve it, sophisticated generator control systems (excitation control with auxiliary signals and fast valving in the generator) are often used. However, when long lines are considered, these control technics are not sufficiently effective to avoid instability. So reactive power compensation is reacted to.

Voltage keeping can be carried out by reactive power control and it is well known that the use of static VAR compensators involves high reactive power control performances. However, maintaining a constant voltage is not sufficient in many cases to obtain a good damping of the power system oscillations; a stabilizing signal is then required.

In this paper, we propose to introduce an auxiliary signal in the SVC voltage regulator, proportional to the transmission line power flow variation. This auxiliary signal could be conceived by the analogical method [1] or by using a modern method as the adaptive control [3]. These two approaches will be compared in different conditions.

I The stabilizing signal

The studied SVC consists of fixed capacitors and thyristor controlled reactors and has been chosen for the rapidity of its response and its possibility to lead and lag reactive power. Within the linear control range, it is assimilated to a variable susceptance B.

The stabilizing signal is proportional to the transmission line power flow variation at the point of connexion of the SVC. A filtering is necessary at the output of the measurement block.

Among the various methods, the self turning control method is the most feasible in terms of implementation and it can be used for no-minimum phase systems. The process is identified by use of an algorithm based on a least square recursive method with a missing factor. The optimal control minimizes a quadratic index performance [2][4]. The block diagram is given on figure 1

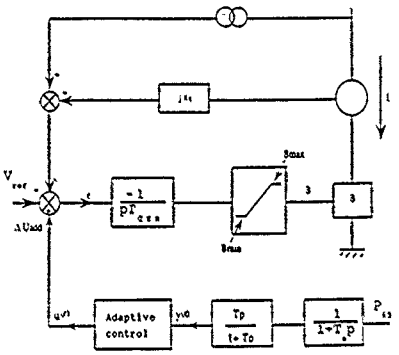


fig 1: Block diagram of the adaptive control.

II Studied power system

It is simply considered a generator feeding a medium power system through a transformer and a long transmission line. The power system is assimilated to an infinite bus bar with a short circuit reactor. Two loads are also plugged in. The generator comprises a synchronous machine with its excitation and governor control. These are chosen without any sophisticated part which can mask the compensator effect. The SVC is connected at the middle of the line.

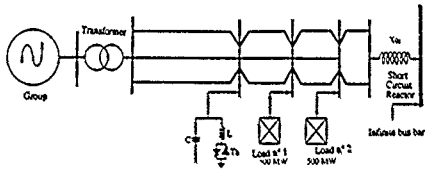


fig 2: The studied power system.

III Test results

Simulations have been done with a multimachine transient stability program developped at Electricité de France.

In case of small disturbances resulting from applying a negative step of 5% on the reference of the generator excitation control, the power system presents an instability. When the SVC is added, the oscillations are slightly damped but it is not sufficient to stabilize the generator.

When the analogical signal is incorporated into the SVC, the oscillations are significantly damped. The curve b on figure 3 has been obtained for a gain of the additional loop equal to 5, but an increase makes the signal less efficient.

The adaptative signal gives better results as it can be seen on curve a. The oscillations are damped rapidly (after 4.5 s). Moreover, since the adaptative signal does not entail an immediate response, which analogical signal does, there is no overvoltage observed at the beginning.

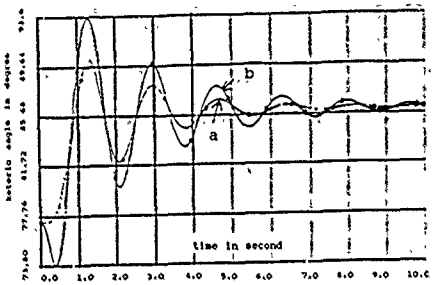


fig 3: Rotoric angle, the SVC is equipped with adaptive (a) or analogical (b) auxiliary signal.

Two simulations of high magnitude disturbances, have been experimented: a reporting load and a three-phase short circuit at the output of the group.

The reporting load consists of the opening of a portion of the line between the group and the point of connexion of the SVC. In this case, the results are similar to those obtained for low disturbances. The oscillations are better damped and the system is more rapidly stabilized with the adaptive signal.

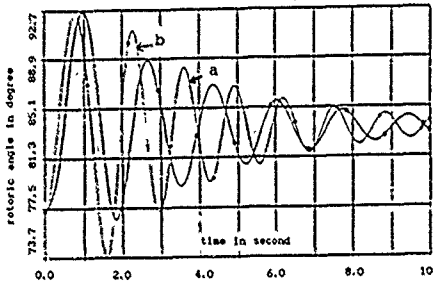


fig 4. Rotoric angle in case of reporting load, the SVC is equipped with adaptive (a) or analogical (b) signal.

In the case of a three-phase short circuit suppressed after 100ms, the results obtained with the adaptive signal, are not satisfactory. A slight damping of the oscillations of the rotoric angle can be observed. Control reaches quickly limit values. This is clear on figure 5b which gives the voltage at the output of the generator. The error obtained at the identification level grows more and more.

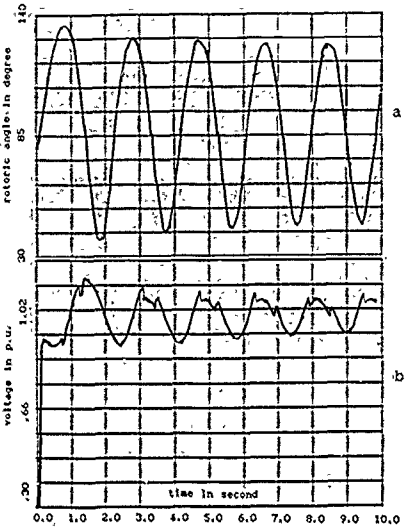


fig 5. Rotoric angle (a) and voltage (b) after a three-phase short circuit, the SVC is equipped with an adaptive signal.

For the same disturbance, the system is stabilized with an analogical signal and the best results are get for a gain equal to 8.

Conclusion

It has been shown in this paper that the SVC with an auxiliary signal proportional to variation of power flow improves dynamic stability. Adaptive control gives much better results than an analogical signal for low magnitude disturbances. Identification still poses a problem in hard conditions and the adopted algorithm needs improving. Another solution would be to fix the parameters during the fault and to use optimal control after a small delay.

Appendix

Generator: S_n=1070 MVA, x_d=x_q=2.47pu, x'_d=.4pu x''_d=.29pu, T'_{do}=8.9s; T''_{do}=.05s; H=2.5s

SVC: X_s=15 kVAR/kV; rating ±300 MVAR; T_{CER}=20ms; T= 50ms; T_s=5ms.

Bibliography

[1] HIROSHIMI KINOSHITA
Improvement of power system dynamic stability by static VAR system. Electrical Engineering Japan vol 99 n°6 1979 pp 81-89

[2] I.D.LANDAU - L.DUGARD
Commande adaptative. Aspects pratiques et théoriques Ed Masson 1986

[3] M.AMOROUAYECHE - C COUNAN - J M KAUFFMANN
Application of adaptive control to static VAR compensator 10° IASTED MIC Innsbrück 1991

[4] M.AMOROUAYECHE
Stabilisation d'un groupe en antenne au moyen d'un compensateur statique. Thèse INPL Nancy - avril 1990

MONTE CARLO SIMULATION OF CURRENT HARMONICS ARISING FROM STATIC POWER CONVERTERS WITH RANDOM FIRING ANGLES

L. Pierrat*,**, Y. J. Wang*, R. Feuillet*

* Laboratoire d'Electrotechnique de Grenoble
CNRS UA-355, BP. 46, Domaine Universitaire
38402 St. Martin d'Hères, France

** Electricité de France
Division Tech Générale, 37 Rue Diderot
38040 Grenoble, France

Abstract -This paper presents a simulation study for the summation of random harmonics produced by a number of independent harmonic generators. The harmonic generators considered are six-pulse converters with random firing angles. At the first stage, a probabilistic converter model is developed to derive the distribution functions of the magnitude and phase of harmonic currents for a single converter. This is followed by considering the combination effect of harmonics of N converters. The statistical behavior of the vectorial sum of N harmonics of the same order is analysed and justified by the results of Monte Carlo simulation. The potential applications of the simulation results are also discussed.

I. INTRODUCTION

Much attention has been given to the problem of random harmonic summation because of the increasing presence and random features of power electronic installations in power systems. In stationary state, the stochastic change in the loads of static converters leads to stochastic variation in the current spectrum. Most publications dealing with random harmonics assume that the magnitude and phase of harmonics are statistically independent, so a phasor can be resolved into two independent components on x-axis and y-axis, and the resultant harmonic can consequently be calculated on each axis [1-4]. Hence, complicated vectorial summation of random variables becomes simple arithmetic summation for each random component. The central limit theorem is often cited to explain the tendency towards normal distribution for each projection. However, neither the assumption of independence between the magnitude and phase of harmonics was assessed, nor the mechanism of harmonic generation of realistic nonlinear loads was considered. In this paper, a type of realistic industrial harmonic generator (i.e., six-pulse converter) is modeled and simulated by Monte Carlo methods to evaluate the validity of the foregoing assumption.

II. CONVERTER MODELS

Six-pulse power converter has been widely used in the industry. Although sophisticated models of this type of converter, that take into account the effect of dc current ripple [5] and commutation [6], are available, the simplest model existing in most standard textbooks (e.g., [7]) is employed in this study. Negligence of commutation and dc current ripple may seem restrictive and unrealistic. It has, however, been studied by simulations that the negligence does not really influence the statistical behavior of resultant harmonics and, moreover, makes possible the analytical formulation of their probability density function (pdf). Assuming that the commutation and ripple effects are neglected and the dc load of the converter is resistive, the magnitude and the phase of the hth harmonic current can be written respectively as (h=1, 5, 7, 11, 13,...):

$$I_h(\alpha) = K_h\cos\alpha; \quad \varphi_h = -h\alpha \qquad (1)$$

where $K_h=18(\sqrt{2})V/(hR\pi^2)$, R is the resistance on dc side, α the firing angle and V the phase voltage in rms value. If the firing angle α varies randomly and uniformly between α_1 and α_2, then the pdf of I_h and φ_h can be obtained by a transformation of random variables [8]:

$$f_h(I_h) = \frac{1}{(\alpha_2-\alpha_1)\sqrt{K_h^2-I_h^2}}, \quad K_h\cos\alpha_2<I_h<K_h\cos\alpha_1 \qquad (2)$$

$$g_h(\varphi_h) = \frac{1}{h\Delta\alpha}, \quad -h\alpha_2<\varphi_h<-h\alpha_1, \Delta\alpha=\alpha_2-\alpha_1 \qquad (3)$$

Eq. (2) and (3) indicate that the harmonic magnitude and phase are statistically independent although α is the only varying parameter that correlates I_h with φ_h. If I_h is expressed in complex form, the physical interpretation of I_h is obvious: I_h is indeed the sum of two random currents:

$$I_h = K_h\cos\alpha[\cos(h\alpha)-j\sin(h\alpha)] = (K_h/2)[1\angle-(h+1)\alpha + 1\angle-(h-1)\alpha] \qquad (4)$$

Each of them has constant magnitude $(K_h/2)$ and random phase $-(h\pm1)\alpha$.

III. RESULTANT HARMONICS OF N CONVERTERS

A computer program based on Monte Carlo methods [10] has been written to simulate the resultant harmonic current of N converters (N=1, 2,...,10). The converters are assumed to have the same power rating and the same variation range of α ($10°<\alpha<70°$) V and R are both set to be equal to 1.0 as base values. All resultant harmonic currents are normalized by a factor of $NI_{h(max)}$, so the values fall in the interval [0,1]. The normalized resultant hth harmonic current is noted as J_h. Note that $I_{h(max)}$ indicates the maximum hth harmonic current generated by one converter.

Fig. 1 illustrates the pdf's of the fundamental current for N= 1, 2, and 10. Fig. 2 shows the pdf's of the 7th harmonic currents for the same N. It is obvious that the pdf's in Fig. 1 and Fig. 2 for N=1 are identical because of the normalization. In fact, for N=1, normalized pdf's are all identical for all h. The analytical formula of $f_h(I_h)$ given in Eq. (2) agrees with the simulation results. Note that the pdf curves are not smooth due to inherent characteristics of Monte Carlo simulations.

When N increases to about 10, the pdf of fundamental resultant current approaches a normal distribution. This result is not surprising since the possible variation range of φ_1 ($=\Delta\alpha$) is confined within (0°, 90°) and the simulation reduces again the range to (10°, 70°) Vectorial summation of such vectors with little freedom of phase variation approaches evidently their arithmetical summation that follows a normal distribution according to the central limit theorem.

On the other hand, the pdf of the 7th harmonic resultant becomes a Rayleigh distribution as N increases [9]. Although not illustrated in the paper, it can be shown that the higher the harmonic order is (from h≥5), the more rapidly the pdf of the resultant harmonic current approaches a Rayleigh distribution. This phenomenon can be explained by the fact that $g_h(\varphi_h)$ becomes more and more uniform in the interval $[-\pi, \pi]$ when the product $h\Delta\alpha$ increases. As long as N is large enough to apply the central limit theorem to the sum of resolved components, the resultant approaches a Rayleigh distribution [2].

It should nevertheless be noted here that $g_h(\varphi_h)$ is not exactly uniform in the interval $[-\pi, \pi]$ although it is uniform in $[-h\alpha_2, -h\alpha_1]$.

As N increases to a certain level, the resultant harmonic deviates from a Rayleigh distribution to a normal distribution. In our simulations, N=10 is not large enough to show the effect of inexact uniform phase but is sufficiently large for the application of the central limit theorem. The problem of imperfect uniform phase distribution is to be discussed in a future paper.

IV. POTENTIAL APPLICATIONS

The prediction of harmonic level is important for many applications such as sizing of harmonic filters and step-down transformers, reactive power compensation, harmonic penetration analysis, .. and so on. It is, however, more desirable to know the statistical distribution of harmonic level than the value given by some other summation critera (i.e., arithmetic sum, root sum square,.. etc.). For example, one may wish to know the value for which D% of the time the resultant harmonic current is smaller than it. D% is often called the non-exceeding probability and is conventionally chosen to be 99%, 97.5% or 95% depending on the requirements of the application.

Fig. 3 illustrates the variation of normalized resultant harmonic current J_h at a non-exceeding probability equal to 99% versus the number of converters N. The physical interpretation of J_h is indeed the harmonic diversity factor. It can be seen that J_h decreases with N. When N=10, the magnitude of the resultant harmonic current of order 5, 7, and 11 has 99% of probability to be equal or less than about 60% of their arithmetic sum.

V. CONCLUSIONS

The physical interpretation of random harmonic currents absorbed by a six-pulse converter is given. The discrepancy between ideal random phasors and random harmonics generated by a realistic nonlinear load can consequently be assessed. In general, Rayleigh distribution is a good approximation of resultant harmonic distribution. The quality of Rayleigh distribution approximation becomes better when $h\Delta\alpha$ increases. Good agreement is obtained between the results of Monte Carlo simulation and analytical formulation. The simulation gives also information on harmonic diversity factor at a designated non-exceeding probability, which is useful for many applications concerning harmonic penetration problems.

VI. REFERENCES

[1] W. G. Sherman, "Summation of harmonics with random phase angles", Proc. IEE, Vol. 119, No. 11, pp. 1643-1648, Nov. 1972

[2] N. B. Rowe, "The summation of randomly varying phasors or vectors with particular reference to harmonic levels", IEE Conf. Publ., No. 110, pp. 177-181, 1974

[3] A. Kloss, "Statistical analysis of harmonic problems in power electronic installations" (in German), Bull. Assoc. Suis Electr., vol. 66, pp. 427-433, April 1975

[4] Y Baghzouz & O. T. Tan, "Probabilistic modeling of power system harmonics", IEEE Trans. on Indust. Appl., Vol. IA-23, No. 1, pp. 173-180, Jan/Feb 1987

[5] L. G. Dobinson, "Closer accord on harmonics", Electronics & Power, pp. 567-572, May 1975

[6] A. D. Graham & E. T. Schonholzer, "Line harmonics of converters with dc motor loads", IEEE Trans. on Indust. Appl., Vol. IA-19, No. 1, pp. 84-93, Jan/Feb 1983

[7] B. K. Bose, "Power Electronics and AC Drives", Prentice-Hall, 1986

[8] A. Papoulis, "Probability Random Variables and Stochastic Processes", McGraw Hill, 1965

[9] L. Pierrat, "Combination of vectors with random phase differences - Quasi-invariance of the resultant amplitude distribution" (in French), to be presented at La 23e journée de statistique, Strasbourg, France, May 1991

[10] R. W. Hamming, "Numerical methods for scientists and engineers", pp. 383-393, McGraw-Hill, 1962

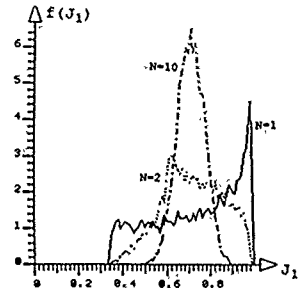


Fig. 1 pdf's of fundamental currents for N=1, 2, and 10

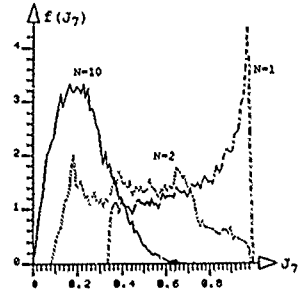


Fig. 2 pdf's of the 7th harmonic currents for N= 1, 2, and 10

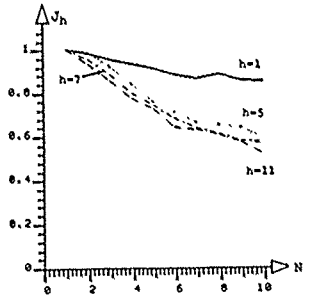


Fig. 3 Variation of J_h versus N (non-exceeding probability =99%)

IMPLEMENTING AN ELECTRONIC TOPOLOGICAL PICTUREBOOK

VAROL AKMAN, AHMET ARSLAN
Bilkent University
Bilkent, 06533 Ankara, Turkey

WM. RANDOLPH FRANKLIN
Rensselaer Polytechnic Institute
Troy, New York 12180-3590, USA

Abstract

An electronic topological picturebook is envisaged as a computerised version of George K. Francis' *A Topological Picturebook*, Springer-Verlag (1987). Francis' book is full of complicated topological figures, mostly drawn manually. Our ultimate goal is to automate the production of such illustrations and to obtain publication-quality hardcopy using assorted techniques of graphics. The present paper can be regarded as a modest first attempt in that direction.

1 INTRODUCTION

Several methods exist for generating computer models of real or imaginary objects. The most popular approach is to use polygons as low level primitives which define more complex objects. Obviously, it is difficult and time consuming for a designer to define an object by such simple, low level primitives. Therefore, a higher level primitive such as a B-spline (or a Bézier) surface is preferable.

In this paper, the suggested techniques for representing 3-D shapes are all based on some kind of sweeping. The implemented program TB (which stands for 'Topology Book') is a rudimentary graphical workbench to help topologists illustrate their ideas more effectively. Central to our implementation is a paradigm of solid modelling, viz. *shape = sweep + control* [13]. TB is written in the C programming language and runs on a colour Sun workstation. Our other papers, covering TB or treating related issues, include [2, 3, 4, 5]. An early yet elegant work which was inspirational for us is Baumgart's geometric editor [7].

This paper owes its existence to *A Topological Picturebook* of George K. Francis—a book that was written to encourage mathematicians to illustrate their work and to help artists to understand the abstract ideas expressed by such drawings. Here we are running the risk of oversimplifying Francis' work considerably for we cannot yet match the quality and the complexity of the figures in that book, viz. the two figures reproduced in [10]—these are known as the 'tetrahedral eight knot' and have to do with Thurston's acclaimed work (cf. [12] for a popular account) in 3-manifolds.

A topological picturebook may sound somewhat futile vis-à-vis the fact that sketching and visual presentation of topological constructions/proofs are slowly losing their stronghold they once held. It may appear that the science of the 'deformation of shapes' is becoming sterile in terms of figures. However, we believe that there is definite place for a topological picturebook of the sort to be described here. The following excerpt succinctly explains our view [10].

> The pedagogy that underlies the entire book, and which I bring out specially here, comes from Bernard Morin of Strasbourg. Pictures without formulas mislead, formulas without pictures confuse. I don't know if Bernard would say it this way, but it is how I have understood his work...Morin's vivid, pictorial description of his bold constructions has inspired their realisation in many a drawing, model, computer graphic and film. But he insists that ultimately, pictorial descriptions should also be clothed in the analytical garb of traditional mathematics."

2 THE SWEEP PARADIGM

A class of solids called *nonprofiled sweep objects* is defined by two parametric curves, a 2-D *contour* and a 3-D *trajectory*. The contour is moved along the trajectory to generate a parametric surface. A classification of sweep objects is given by Bronsvoort et al. [8].

- Translational sweep: The contour is arbitrary and the trajectory is a straight line.
- Rotational sweep. The contour is arbitrary and the trajectory is a circle.
- Circle sweep. The contour is circular and the trajectory is arbitrary.
- General sweep: Both the contour and the trajectory are arbitrary.

A nonprofiled generalised cylinder is defined by an arbitrary closed 2-D contour and an arbitrary 3-D trajectory [8, 15]. Here, the 2-D contour c can be defined in the parametric form as $c(v) = (c_x(v), c_y(v))$, where $v_i \le v \le v_f$ and $c(v_i) = c(v_f)$. As a parameter, v varies from v_i to v_f. The parametric functions c_x and c_y trace out the contour. The 3-D trajectory t can be defined in the parametric form as $t(u) = (t_x(u), t_y(u), t_z(u))$, where $u_i \le u \le u_f$. As a parameter, u varies from u_i to u_f. The parametric functions t_x, t_y and t_z trace out the trajectory. For a nonprofiled generalised cylinder, the contour c moves along the trajectory t.

The spatial relation between the contour and the trajectory must be defined at every point of the trajectory. The contour plane is perpendicular to e_1, the unit vector tangent to the trajectory. As e_3, a fixed unit normal to the plane of the trajectory is chosen. Note that e_2 is the vector product of e_3 and e_1. Consequently, it is possible to present a nonprofiled generalised cylinder as a 5-D vector function T_{np} of two parameters u, v as

$$T_{np}(u, v) = (t_x(u), t_y(u), t_z(u), c_x(v) \cdot e_2(u), c_y(v) \cdot e_3(u)),$$

where $u_i \le u \le u_f$ and $v_i \le v \le v_f$. Here, there is a problem if the curvature of the trajectory is zero. In this case, some approximations are necessary. For example, the direction of the tangent vectors for the first and the last points of the trajectory can be selected as directions of the first and the last segments. For other points, the direction of the vectors as specified by the previous and the next points can be selected.

If some deformations are necessary on a nonprofiled generalised cylinder by scaling, then a profiled generalised cylinder can be defined as a mathematical model [8, 15]. The contour can be scaled independently in x and y directions, and is determined by e_2 and e_3. For each point of the trajectory, two scale factors S_x and S_y are specified. The scale functions are expressed in terms of the parameter of the trajectory: $S(u) = (S_x(u), S_y(u))$, where $u_i \le u \le u_f$. If the scale factor is substituted in T_{np} defined in the previous section, then it will have an effect only on the contour functions, and the resultant profiled generalised cylinder will be defined as a 5-D vector function T_p.

$$T_p(u, v) = (t_x(u), t_y(u), t_z(u), S_x(u)\ c_x(v)\ e_2(u), S_y(u)\ c_y(v)\ e_3(u)),$$

where $u_i \le u \le u_f$ and $v_i \le v \le v_f$.

3 NORMAL FORMS FOR SURFACES

Classification of 2-manifolds is a completely solved problem of topology. Here we omit classical introductions to the subject (cf. Ringel [16] for that) and just summarise the results.

The *plane representation of a polyhedron* is all the information provided by the polygons in the plane representation, the pairing of the sides and the orientation of the sides (before identification).

Given a polyhedron P and its symbolic representation scheme Σ, one defines four *elementary operations*. The property of these operations is that although they do change P as well as Σ, the surface itself is not changed. The elementary operations are as follows:

- SU1 (Subdivision of dimension one): An edge of the polyhedron is divided into two new edges by taking an inner point of the edge as an additional vertex.
- CO1 (Composition of dimension one): This is the reverse operation of SU1.
- SU2 (Subdivision of dimension two): Two vertices of a polygon in the polyhedron will be connected by a new edge dividing the polygon into two new polygons.
- CO2 (Composition of dimension two): This is the reverse operation of SU2.

Two polyhedra P and P' are said to be *elementarily related* if P can be transformed into P' by using a finite number of the elementary operations SU1, CO1, SU2 and CO2. A polyhedron is called *orientable* if one can choose an orientation for each polygon such that each edge is used in both possible directions. It turns out, by the fundamental theorem of the classification of 2-manifolds, that each polyhedron is elementarily related to one of the following normal forms:

$$(H_0) \quad ea-$$

$$(H_p) \quad a_1b_1a_1-b_1-\ldots a_pb_pa_p-b_p-$$

$$(C_q) \quad c_1c_1c_2c_2\ldots c_qc_q$$

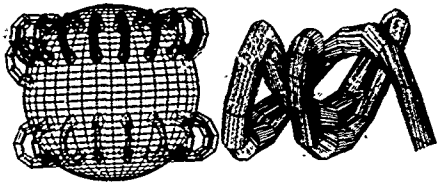


Fig. 1 A k-tori and a knot drawn by our program

4 REALISM

There are assorted techniques that can be used to give a realistic view of an object as a 2-D image. The following are some important ones taken from Francis [10]:

1. 'For complicated objects it is often impossible to find a view which does not hide some important structure behind a surface sheet. One remedy is to remove a regular patch from the object, creating a transparent *window* through which this structure can be seen in the picture.'
2. 'To distinguish an edge whose other face is hidden from an edge which merely separates two visible faces, drafting teachers recommend heavier lines for the former relative to the latter.'
3. 'The shading technique I use makes no pretense of accuracy and realism. It merely encodes positional information and helps distinguish rounded contours from sharp borders. It is based on a few optical principles.'
4. 'Line patterns based on boxes in affine projection suffer from the Necker cube illusion: 'Whic_ is front and which is back?' Thickening the facing borders helps decide which view is intended.'

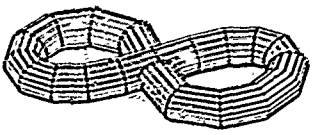


Fig. 2 A torus drawn by our program

5 CONCLUSION

The interaction of the TD with the user is straightforward. For example, a torus can be rendered by selecting only five control points to produce the contours and a radius. Our fundamental aim is to hide details of the processing from the user. In other words, the user will design objects as if he or she is doing free-form sculpturing, by carrying out high level operations. This is in the precise spirit of Pentland's sketching system [14].

References

[1] Abelson, H., and diSessa, A., *Turtle Geometry: The Computer as a Medium for Exploring Mathematics*, MIT Press (1982).

[2] Akman, V., *Steps into a Geometer's Workbench*, Report CS-R8726, Centre for Mathematics and Computer Science, Amsterdam (1987).

[3] Akman, V., *Unobstructed Shortest Paths in Polyhedral Environments*, Lecture Notes in CS, Vol. 251, Springer-Verlag (1987).

[4] Akman, V., and Arslan, A., 'An electronic topological picturebook,' *NATO ASI on Cognitive and Linguistic Aspects of Geographic Space*, Las Navas del Marques, Spain (July 1990).

[5] Arslan, V., Isler, V., and Akman, V., 'A procedure to sweep arbitrary curves,' *Fifth International Symposium on Computer and Information Sciences*, Cappadocia, Turkey (November 1990).

[6] Bartels, R. H., Beatty, J. C., and Barsky, B. A., *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Morgan Kaufmann (1988).

[7] Baumgart, B. G., *Geomed: Geometric Editor*, Report STAN-CS-74-414, CS Department, Stanford University, Stanford, CA (1974).

[8] Bronsvoort, W. F., Nieuwenhuizen, P. R. V., and Post, F. H., 'Display of profiled sweep objects,' *Visual Computer*, Vol. 5, 147-157 (1989).

[9] Cardelli, L., *Building User Interfaces by Direct Manipulation*, Report 22, System Research Centre, Digital Equipment Corporation, Palo Alto, CA (1987).

[10] Francis, G. K., *A Topological Picturebook*, Springer-Verlag (1987)

[11] Kauffman, L. H., *On Knots*, Princeton University Press (1987).

[12] Kneale, D., 'Shaping ideas: A topologist wows the world of math by seeing the unseen,' *Wall Street Journal* (March 18, 1983).

[13] Martin, R. R., and Stepson, P. C., 'Sweeping of three-dimensional objects,' *CAD*, Vol 22, No. 4, 223-234 (1990).

[14] Pentland, A. P., *Supersketch™ User Manual*, AI Centre, SRI International, Menlo Park, CA (1985).

[15] Post, P. H., and Klok, F., 'Deformations of sweep objects in solid modeling,' Requicha, A. A. G. (ed.), *Eurographics'86 Proceedings*, North-Holland, 103-114 (1986)

[16] Ringel, G., *Map Color Theorem*, Springer-Verlag (1974)

MULTI-TIME SCALE DECOMPOSITION OF THE SYNCHRONOUS MACHINE MODEL

H. GUESBAOUI ; O. TOUHAMI ; C. JUNG

CRAN/ENSEM - 2 Avenue de la Forêt de Haye
54516 VANDOEUVRE LES NANCY CEDEX
FRANCE

KEYWORDS :

Singular perturbations, model reduction, time scale decomposition, synchronous machine, sub-subtransient behaviour.

Introduction :

The modelling of the variable speed drives is so complex that it makes their analysis difficult. This complexity is all the greater as the machine is an alternative one. In the case of synchronous machines, or of turbogenerators, the model is non linear and highly ordered.
Most studies use linear models and neglect the fast transients, which do not modify the working of the machine as a whole. These simplifications are based on the knowledge of the physical phenomenon and not on an accurate model analysis. Actually, it can be shown that these approaches are agregation ones. So they present the same drawbacks: the informations on the neglected modes cannot be taken into account.
Another difficulty is to evaluate the sensitivity of the errors on the measured parameters toward the simplified model. This is quite impossible to achieve when the simplified model is not obtained by a rigourous mathematical technique.
We propose a multitime scale approach to separate the model into reduced subsystems, which allows to consider the fast behaviours, ie transient and subtransient operations, in the case of a synchronous machine.

1 - Simplification methods :

The equations of a synchronous machine using Park's transformation can be represented by the field winding and the nd and nq damper windings in a set of d-q axis bound to the rotor (fig. 1a). The model is then given by two d-q equivalent circuits, (fig 1b). The stator (la, ra) and the field excitation (l_f, r_f) are separated by the damper windings.
The most commonly used simplification neglects the stator resistance ($r_a = 0$), i.e. supposes that the stator magnetic flux ψ_d and ψ_q reach instantaneously the steady state.
A structural simplification consists in using only one damper winding for the d-axis and two for the q-axis (nd = 1, nq = 2) : in this case, the model is 2x2 with the time constants :

(T"d, T'q) and (T'd, T'q) : respectively the subtransient and transient short-circuit time constants,
(T"d_0, T"q_0) and (T'd_0, T'q_0) : respectively the subtransient and transient open circuit time constants.
A plotting of the frequency characteric responses of 1/Xd (js) and 1/Xq (js) can be computed by :

$$\frac{1}{x_d(js)} = \frac{1}{x_d} \cdot \frac{(1+js\,T'd_0)(1+js\,T''d_0)}{(1+js\,T'd)(1+js\,T''d)} \quad \text{and}$$

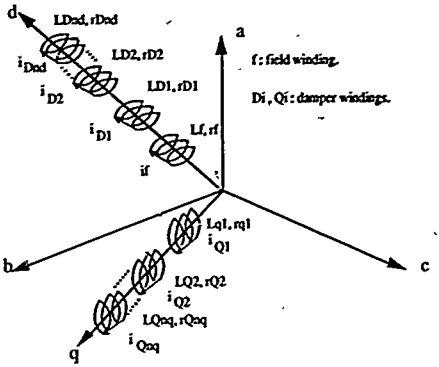


Fig 1a.(d-q)-Park's model.

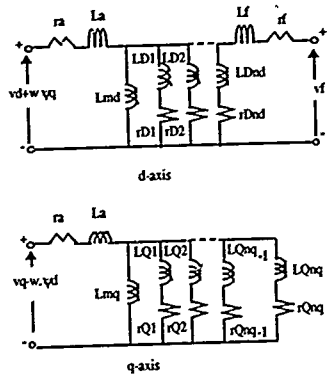
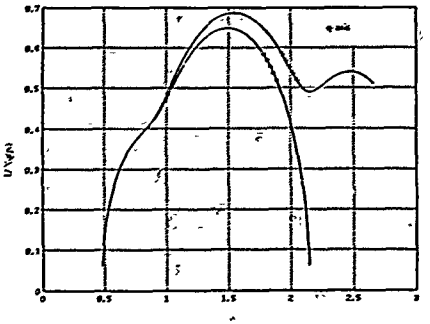


Fig 1b. Equivalents circuits.

$$\frac{1}{x_q(js)} = \frac{1}{x_q} \frac{(1+js\,T'q_0)(1+js\,T''q_0)}{(1+js\,T'q)(1+js\,T''q)}$$

using model [1]. This plotting exhibits two distinct behaviours : transient behaviour and established behaviour, (fig.2). Nethertheless the equations of the model are corrected by d-axis damper winding flux. An additional hypothesis ($r_{D1} = 0$), [2] valid for the solid-iron rotor, allows to reduce the model again.



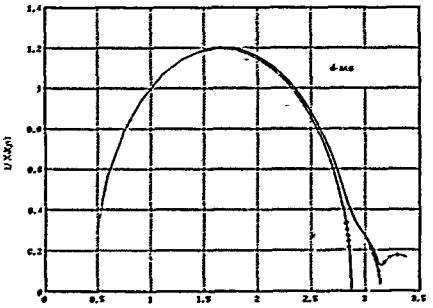


Fig 2. Computational Locus curves of U'Xd(s) and U'Xq(s).

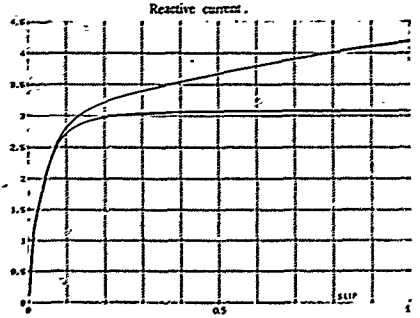
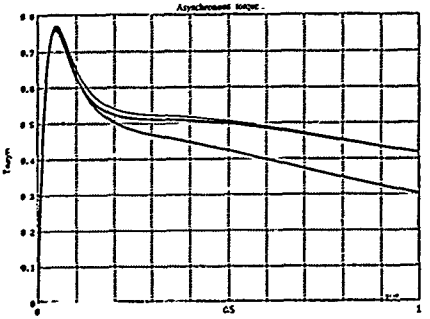


fig3. External characteristics curves.



d-axis q axis

model 2*2 ...
Xd =2,18 Xq=2,112
X'd=0,3565 X"d=0,3479 X'q=0,9872 X"q=0,4652
T'd=1,2499 T"d=0,5956 T'q=1,0173 T"q=0,0998
T'd0=7,644 T"d0=0,6101 T'q0=2,1686 T"q0=0,2117

model 2*3 ...
Xd=2,18 Xq=2,112
X'd=0,3565 X"d=0,3479 X"'d=0,3185 X'q=0,9872 X"q=0 4652 X"'q=0,3277
T'd=1,2499 T"d=0,5956 T"'d=0,0616 T'q=1,0173 T"q=0,0998 T"'q=0,0061
T'd0=7,644 T"d0=0,6101 T"'d0=0,0673 T'q0=2,1686 T"q0=3,2117 T"'q0=0,0086

model 2*4 ...
Xd=2,18
X'd=0,3565 X"d=0,3479 X"'d=0,3185 X""d=0, 2951
T'd=1,2499 T"d=0,5956 T"'d=0,0616 T""d=0,0021
T'd0=7,64 T"d0=0,5956 T"'d0=0,0673 T""d0=0 0023

For a greater number of damper windings (nd = 2, nq = 3 : model 2 x 3) some authors [3] have measured new time constants T"'d and T"'q smaller than T"d and T"q. We have a larger number of observed dynamics with different time-constants. The computational locus curves are similar to the experimental curves of [3] . The new behaviours are called sub-subtransient with the time constants (T"'d, T"'q) and (T""d, T""q) and so on.
We have validated the existence of these different dynamics for a model 2 x 4, (fig. 2).

For a salient pole machine,the asynchronous-torque and the reactive-current characteristics obtained from the sum (1/Xd(js) + 1/Xq(js)) are plotted in fig 3 for the three equivalent circuits (2x2, 2x3 and 2x4). The asynchronous starting emphasies the importance of the part played by the number of damper windings considered in the model. The large number of them shows that the operation of a synchronous machine can be compared to an induction machine with a squirrel-cage winding.

All these simplifications are in fact singular perturbation hypothesis.

2 - Multitime scale technique :

The mechanical and electrical parts are decoupled using the stator synchronous speed ω_s.
The choice of a small parameter multiplying the derivatives of the fast electrical parts is : $\varepsilon = 1 / \omega_s$, [4].

The electrical equations are :

$$\begin{cases} \frac{d}{dt} \underline{Y} = - A(\omega) . \underline{Y} + \underline{v} \\ A(\omega) = r.L^{-1} + \omega.J^* \end{cases}$$

where : ω is the rotor speed
$\underline{\Psi}$ and $\underline{v}$ are respectively the flux and the voltage vectors given by
$\underline{\Psi} = (\Psi_f, \Psi_{D1}, \Psi_{D2}, \ldots, \Psi_{Dnd}, \Psi_{Q1}, \Psi_{Q2}, \ldots\ldots, \Psi_{Qnq}, \Psi_d, \Psi_q)^T$
$\underline{v} = (V_f, 0, 0, \ldots\ldots\ldots\ldots 0,0,\ldots\ldots, 0,0, \ldots\ldots\ldots\ldots\ldots, v_d, v_q)^T$

and the matrix resistances:

$\underline{r} = (r_f, r_{D1}, r_{D2}, \ldots r_{Dnd}, r_{Q1}, r_{Q2}, \ldots, r_{Qnq}, r_a, r_a).I_{nd+nq+3}$,
where :

$I_{nd+nq+3}$ is a $(nd + nq + 3) \times (nd + nq + 3)$ identity matrix.

$A(\omega)$, L and J^* are $(nd + nq + 3) \times (nd + nq + 3)$ matrices.

$$L = \begin{pmatrix} L_f & M \\ M^T & L_{dq} \end{pmatrix} \qquad J^* = \begin{pmatrix} 0 & 0 \\ 0 & J \end{pmatrix}$$

with 2×2 matrices: $L_{d,q} = \begin{pmatrix} L_d & 0 \\ 0 & L_q \end{pmatrix}$ and $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

L_f : $(nd + nq + 1) \times (nd + nq + 1)$ self rotor matrix ,
M : $(nd + nq + 1) \times 2$ mutual self inductance.

For the case nd = nq = 1, the eigenvalues of A (ω) are a set of three groups, [5] :

$$\lambda_{1,2} \approx -\frac{1}{\tau''_H} \pm j\omega \qquad \frac{1}{\tau''_H} = \frac{r_a}{2}\left(\frac{1}{L''d} + \frac{1}{L''q}\right)$$

τ''_H is the armature time constant

$$\begin{cases} \lambda_3 \approx -\dfrac{1}{T_{D_1}} = -\dfrac{1}{T''d} \\ \lambda_4 \approx -\dfrac{1}{T_{Q1}} = -\dfrac{1}{T''q} \end{cases}$$

and

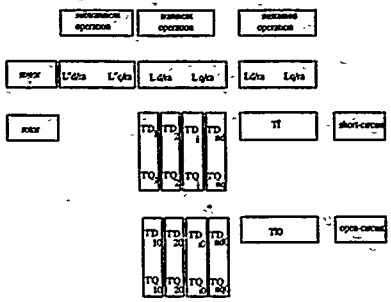
$$\lambda_5 \approx -\frac{1}{T'_f} = -\frac{r_f}{L'_f}$$

with : $L'f = \sigma_{df}.Lf$ and $Tf = L'f/rf \approx Td$.

The flux advance with three time scales :
(Ψ_d, Ψ_q) : substransient flux with $\tau_1 = \omega_s t$,
(Ψ_{D1}, Ψ_{Q1}) : transient flux with $\tau_2 = \dfrac{t}{T_{D1}}$ and
Ψ_f : established flux with $\tau_3 = t$.

For the case nd = nq = 2, four time scales must be used And the damper windings flux (Ψ_{D1}, Ψ_{Q1}) and (Ψ_{D2}, Ψ_{Q2}) are transient quantities depending on the reduced times $\tau_2 = t/T_{D1}$ and $\tau_3 = t/T_{D2}$.
The former time constants T'''_d and T'''_q are defined by :
$T_{D2} = T'''_d$, $T_{Q2} = T'''_q$ and $T_{D20} = T'''_{d0}$, $T_{Q20} = T'''_{q0}$.

For the cases nd and nq > 1, the different operations can be summarized on the following scheme :



The stator fluxes are the fast modes and so are the subtransient variables. The expression of the quantities L"d and L"q depends on the nd and nq number of damper windings .The rotor operation is slower and describes the transient dynamics .

Conclusion :

The singular perturbation reduction of the model sets the different behaviours of the synchronous machine in the right order. The sub-subtransient modes are in fact transient modes. The damper windings contribute to these dynamics. They are to be taken into account in the case of a synchronous machine supplied for example by a static converter where the observed modes are very fast.

References :

[1] KAMABU T. and MAUN JC. (1987) : Turbine generator models by the finite element method - IMACS - TC1 - Laval - Quebec - Canada

[2] ANDERSON PM. and FOUAD A. (1980) : Power system control and stability 2nd Ed. Iowa - Vol. CAS - 29 (11) - 782.

[3] CANAY IM. (1988) Physical significance of sub-subtransient quantities in dynamic .behaviour of synchronous machines - IEE Proc., Vol. 135, n° 6, Nov. 1988.

[4] KOKOTOVIC PV. AND SAUER PW. (1989) : Integral manifold as a tool for reduced-order modeling of non linear systems : A synchronous machine case study. IEEE Trans. on A.C., Vol. 36, n° 3, March 1989.

[5] GUESBAOUI H. and IUNG C. (1988) . Simplification of the model of a current fed controlled synchronous machine by the multitime scales method.12th IMACS World Congress - Paris - July 1988.

ABNORMAL MODES OF OPERATION IN DIGITAL FILTERS: A COMPUTER-ASSISTED STUDY

Stanisław Mitkowski and Maciej J.Ogorzalek
IMISUE
Department of Electrical Engineering
Academy of Mining and Metallurgy
al.Mickiewicza 30
30-059 Krakow , Poland

Abstract - Using recently developed software package for unconventional digital filter analysis we study undesired modes of operation in second-order digital filters realised in direct form. Apart from many kinds of parasitic overflow oscillations we confirmed by computer experiments the existence of several intersting types of behavior. These include chaotic oscillations and fractal patterns of trajectories in the filter with modular arithmetic and devil's staircase changes in periods of oscillations in the filter with saturation arithmetic. Some new bifurcation phenomena like "fan-type" bifurcation sequences have also been found.

Introduction

Abnormal modes of operation in signal processing circuits, both analog and digital, for a long time attracted attention of scientists. In digital systems the problems of quantisation effects and overflow oscillations have been analysed by many researchers (see eg [1], [6], [8]) Chua and Lin [2] reported on aperiodic (chaotic) oscillations from second order digital filter employing 2's complement (modulo) arithmetic and operating at the border of the stability domain. Successively in the paper [4] we studied bifurcation phenomena in a digital filter with saturation-type adder overflow characteristic. For the sake of in-depth study of these phenomena we developed a set of specific computer programs implementing newly developed algorithms [5] for simulation of second and third order filters with possibilities of choosing the way of internal coding of signals and type of arithmetic used. The following types of analysis have been implemented so far :

1. Time evolution of system's responses,
2. State-space plots including possibilities of observation of fractal trajectory structure,
3. Investigations of limit cycles :
   - computation of periods of orbits,
   - computation of winding numbers of orbits
   - computation of basins of attraction of periodic orbits,
4. Computation of one-parameter bifurcation diagrams,
5. Construction of winding number diagrams ("devil's staircase")

Below we present chosen results of our simulation studies in second order digital filters whose dynamics are described by a state equation of the form (1-2) In particular our interest is concentrated on nonlinear effects caused by the adder overflow nonlinearity.

$$x_1(k+1) = x_2(k) \quad (1)$$
$$x_2(k+1) = F[bx_1(k) + ax_2(k)] \quad (2)$$

Where : $a, b \in R$,
$FR \to R$, $F(\sigma) = \sigma$ for $|\sigma| < 1$, $F(\sigma) = 1$ for $\sigma \geq 1$, $F(\sigma) = -1$ for $\sigma \leq -1$ (saturation arithmetic) or $F(\sigma) = \sigma$ for $|\sigma| < 1$, $F(\sigma) = \sigma - 2$ for $\sigma \geq 1$, $F(\sigma) = \sigma + 2$ for $\sigma \leq -1$ (modular or 2's complement arithmetic).

Fractal structure of trajectories in the filter with modular arithmetic

Among very interesting dynamic behaviours encountered in the filter with modular arithmetic the most interesting is the chaotic motion observed for $b = -1$ and various a parameters [2]. In Figure 1 we present a trajectory observed in our simulation experiments for $a = 0.5$ and initial conditions $x_1 = -.6135, x_2 = -x_1$, 30000 iterations. It has self-similar structure - repeating, diminishing empty sets of elipse-shape are clearly visible.

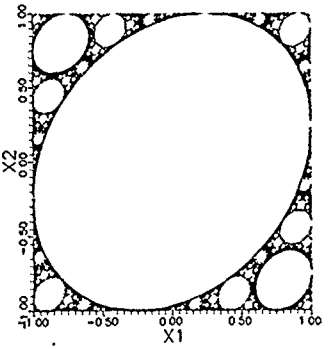


Fig. 1: Chaotic trajectory of a second order digital filter with modular overflow characteristic for $a = 0.5, b = -1$. Note the fractal self-similar structure of the limit set.

Complex bifurcation sequences in the filter with saturation arithmetic

Using our simulation package we were able to discover several new interesting phenomena in the case of a filter with saturation-type adder overflow characteristic. First of all we discovered very complex bifurcation sequences for the filter operating for some parameter choices outside the linear stability sector [4]. One of such bifurcation sequences is shown in Fig.2.

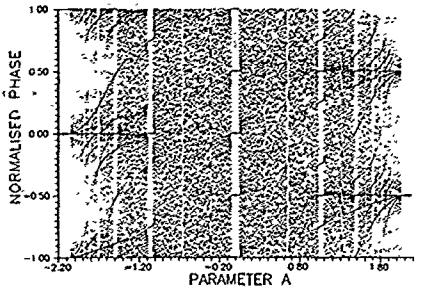


Fig 2. Typical bifurcation sequence observed in the digital filter with saturation arithmetic for the parameter choice outside the stability sector.

It is worthwhile pointing out the existence of extremely complex orbits (probably nonperiodic) and unusual bifurcation sequences which we called "fan-type" shown in the magnified picture in Fig 3.

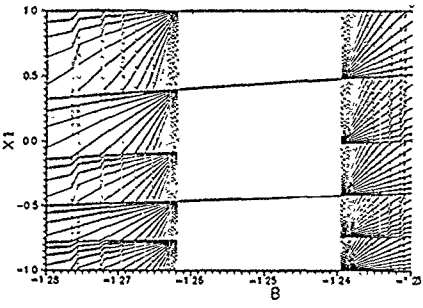


Fig. 3. New "fan-type" bifurcation sequence observed in the digital filter with saturation arithmetic.

For some parameter ranges we discovered an abundance of periodic orbits of all periods depending on the parameter value chosen - the filter reproduces so-called devil's staircase structure of changes of orbit winding numbers when changing the bifurcation parameter - see eg. Fig.4.

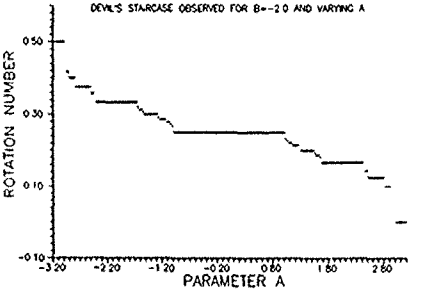


Fig. 4: Complete devil's staircase structure of winding number changes observed in the digital filter with saturation arithmetic.

An extremely interesting result of our study of nonlinear dynamics of digital filters is the fractal structure of basins of attraction of some periodic orbits. For some parameter choices the set of initial conditions for orbits appoaching even simple periodic orbits (as we found eg. for a period four orbit) has fractal structure - its area is finite but the length of its boundary tends to infinity.

Conclusions

Our simulation study revealed extremely interesting phenomena associated with inherent nonlinearities within the digital filter - namely its adder overflow characteristic. The following we found most interesting and requireing further study :

- There exist extremely complex (chaotic) trajectories both in the case of modular and saturation arithmetic,
- Fractal, self-similar structures of limit sets of trajectories are often encountered,
- Existence of very complex periodic orbit structure with devil's staircase sequence of changes of winding numbers has been confirmed,
- New bifurcation structures eg. "fan-type" have been found,
- Several cases of fractal boundaries of basins of attraction of periodic orbits were found in computer experiments.

References

[1] H.J.Butterweck, J.Ritzerfeld, M.Werter " Finite wordlength effects in digital filters". AEÜ, No.2, pp.76-89, 1989.

[2] L.O.Chua, T.Lin "Chaos in Digital Filters". IEEE Trans. CAS, CAS-35, pp 648-658, 1988.

[3] K Falconer "Fractal Geometry. Mathematical Foundations and Applications". John Wiley & Sons , 1990.

[4] Z.Galias, M.J.Ogorzalek "Bifurcation Phenomena in Second Order Digital Filter with Saturation-Type Adder Overflow Characteristic" IEEE Trans. CAS, CAS-37, No.8, pp.1068-1070, 1990.

[5] Z Galias, S.Mitkowski, M.J.Ogorzalek "KRAKFIL - a software toolkit for unconventional digital filter analysis". Submitted for presentation at the European Conference on Circuit Theory and Design, ECCTD'91, Helsingor.

[6] D.Mitra "Large Amplitude Self-sustained Oscillations in Difference Equations which Describe Digital Filter Sections Using Saturation Arithmetic". IEEE Trans. ASSP, ASSP-25, pp.134-143, 1977.

[7] M.J.Ogorzalek, Z.Galias "Arnold tongues and devil's staircase in a digital filter with saturation arithmetic". (submitted to IEEE Trans. CAS).

[8] A.N.Wilson "Limit Cycles Due to Adder Overflow in Digital Filters". IEEE Trans. CT, CT-19, No 4, pp 312-345, 1972.

AN EFFICIENT IMPLEMENTATION OF VARIABLE -INTERCHANGE METHOD FOR THE SOLUTION OF AN ELECTROCHEMICAL MACHINING PROBLEM

TURGUT ÖZİŞ
İnönü University, Department of
Mathematics, Campus, Malatya 44069, TURKEY

Abstract- A two dimensional electrochemical machining problem which consist of a circular anode placed inside a circular cathode is solved using a variable-interchange method. The notching of an initially circular anode due to circular cathode is studied by solving the resulting problem numerically.

1. INTRODUCTION

There are many real-world problems in which the governing differential equation is known but thesolution domain is not completely specified. The term free-boundary problem is commonly used when the boundary is stationary and steady-state problem exists. Moving boundries are associated with time-dependent problems and the position of the boundary has to be determined as a function of time and space. There are, however, some important problems in which the boundary is moving but the equation is elliptic, i.e. they are degenerate problems. Examples are provided by problems in electrochemical machining and the Hale-Shaw flow associated with the injection of fluid into a narrow channel. Therefore, the following discussion is than given in the context of degenerate problems. Electroc· ·ical machining is a technological process in which a workpiece is placed as the anode in an electrolytic cell with a properly shaped cathodic tool so that a desired shape of cathodic work piece is obtained by the electrochemical process.
Hougaard[6] gave references to pioneer papers and determined the stationary anode profile using complex variable methods. Chritiansen and Rasmussen[2] developed a method in which the potential problem was formulated as an integral equation of the first kind. Later, Hansen and Holm[5] used an integral equation of the second kind for similar problem. Meyer [7] developed a method based on the method of lines. Crowley [3] used enthalpy type formulation and Elliott [4] proposed an elliptic variational inequality formulation for this problem.

In this note, single variable-interchange method used for the annular electrochemical machining problem by the aid of Boadway's[1] transformation and this implementation is more economical computationally on this specific problem in comparison to its predecessors.

2. FORMULATION

An approxmate quasi-steady model for the process consist of a boundary value problem for the potential between the electrodes and an equation relating the change of the anode surface to the normal gradient of potential at the anode. Hence, the mathematical formulation is

$$r^2\frac{\partial^2 u}{\partial r^2}+r\frac{\partial u}{\partial r}+\frac{\partial^2 u}{\partial \theta^2}=0 \quad \text{in } D \qquad (1)$$

with boundary conditions

$$u=-v<0 \qquad (2.a)$$

on the cathode $r=a$ and

$$u=0 \qquad (2.b)$$

at the anode $s(r,t)=0$.
The second condition which governs the change with respect to time of the anode surface is:

$$M\frac{\partial s}{\partial t}=-\nabla u.\nabla s \quad \text{on } s(r,t)=0 . \qquad (2.c)$$

Thus, we have a one-phase problem with an elliptic type equation and Stefan type boundary condition with non-zero latent heat.

3. SOLUTION PROCEDURE

For geometries of the problems which involve circular boundaries, it is more convenient to work in polar coordinates. Therefore, we require the solution of the equation(1) in D bounded by $r=a$, on which $u=-v$ and by moving interface $r=s(\theta,t)$ on which $u=0$ for $0\leq\theta\leq\pi/2$
The additional conditions are:

$$u_\theta=0 \quad \text{at } \theta=0 \text{ and } \pi/2. \qquad (3)$$

The single variable-interchange(see,Boadway [1]) of $u(r,\theta)$ to $r(u,\theta)$ renders the equation (1) to the following form:

$$r^2\left(-\frac{\partial^2 r}{\partial u}\right)+r\left(\frac{\partial r}{\partial u}\right)^2+\left[\frac{\partial r}{\partial u}\left(\frac{\partial r}{\partial \theta}\cdot\frac{\partial^2 r}{\partial u\partial\theta}-\frac{\partial r}{\partial u}\cdot\right.\right.$$

$$\left.\left.\frac{\partial^2 r}{\partial \theta^2}\right)-\frac{\partial r}{\partial \theta}\left(\frac{\partial r}{\partial \theta}\cdot\frac{\partial^2 r}{\partial u^2}-\frac{\partial r}{\partial u}\cdot\frac{\partial^2 r}{\partial \theta\partial u}\right)\right] \qquad (4)$$

However, themoving interface condition on the anode must be transformed into an expression for the speed of tho anode along each ray. Therefore, if r be the non-dimensional position vector for the anode surface, then,

$$\frac{\partial r}{\partial t}=\frac{\partial s}{\partial t}+\frac{\partial s}{\partial \theta}\cdot\frac{\partial \theta}{\partial t}=\frac{\partial s}{\partial t}+\frac{\partial s}{\partial \theta}\cdot\frac{1}{s^2}\cdot\frac{\partial u}{\partial \theta}.$$

It is convenient to repace $\partial u/\partial\theta$ in terms of $\partial u/\partial r$. Since, the tangential derivative vanishes on the anode, we can write

$$\partial u/\partial\theta=-(\partial s/\partial\theta)(\partial u/\partial r)$$

So that the gradient condition on the surface lead to

$$\frac{\partial s}{\partial t} = \left(1 + \left(\frac{1}{s} \cdot \frac{\partial s}{\partial \theta}\right)^2\right) \frac{\partial u}{\partial r} \quad . \qquad (5)$$

Discretization of equations: For numerical solution, the transformed potential equation (4) may be discretized and rearranged to yield,

$$A(r_{ij}^n)\left(r_{ij}^{n+1}\right)^2 + B(r_{ij}^n) r_{ij}^{n+1} + C(r_{ij}^n) = 0 \qquad (6)$$

where

$$A(r_{ij}^n) = -\frac{r_{i,j}^n - 2r_{ij}^n + r_{i-1,j}^n}{(\delta u)^2}$$

$$B(r_{ij}^n) = \frac{r_{i,j}^n - r_{i-1,j}^n}{2.\delta u}$$

$$C(r_{ij}^n) = \frac{r_{i,j}^n - r_{i-1,j}^n}{2.\delta u}\left(\frac{r_{i,j+1}^n - r_{i,j-1}^n}{2\delta\theta}\right) \cdot \frac{r_{i-1,j+1}^n - r_{i-1,j-1}^n + r_{i,j+1}^n - r_{i,j-1}^n}{4.\delta u.\delta\theta} - \frac{r_{i,j}^n - r_{i-1,j}^n}{2\delta u} \cdot$$

$$\frac{r_{i,j+1}^n - 2r_{ij}^n + r_{i,j-1}^n}{(\delta\theta)^2}\Bigg) - \frac{r_{i,j+1}^n - r_{i,j-1}^n}{2\delta\theta}\left(\frac{r_{i,j+1}^n - r_{i,j-1}^n}{2\delta\theta} \frac{r_{i,j}^n - 2r_{ij}^n + r_{i-1,j}^n}{(\delta u)^2}\right.$$

$$\left.\frac{r_{i,j}^n - r_{i-1,j}^n}{2.\delta u} \cdot \frac{r_{i-1,j+1}^n - r_{i-1,j-1}^n - r_{i,j+1}^n + r_{i,j-1}^n}{4.\delta\theta\,\delta u}\right)$$

and r_{ij}^n is the nth iteration step value of r_{ij}. The moving boundary condition given by equation(5) can also be discretized expilicitly to give

$$r_{Nj}^{k+1} - r_{Nj}^k = \delta t\left(1 + \frac{1}{(r_{Nj}^k)^2}\left(\frac{r_{N,j+1}^k - r_{N,j-1}^k}{2.\delta\theta}\right)^2\right) \frac{\delta u}{r_{Nj}^k - r_{N-1,j}^k} \qquad (7)$$

where r_{Nj}^k is the kth time step value of r_{Nj}. The moving boundary equation(7), however can easily be solved by an explicit method. But, the equation(6), if the values of r_{ij}^n are all known in nth iteration step, for example, may be solved by regula-falsi method for next iteration step.

Algorithm: To complate the numerical procedure, the test problem can easily be solved by the algorithm stated below:

a) solve the clasical mixed boundary value problem given by equation(6) and relavent boundary conditions as the anode treated as a fixed surface,

b) use equation(7) to determine a new anode surface which becomes the fixed boundary in the next time step,

c) go to step (a).

4. NUMERICAL EXAMPLE

To show the applicability of the method, let us consider the notching of an initially circular anode due to circular cathode. In this problem, it is assumed that the anode surface is insulated everywhere, except for $\theta\in(3\pi/8, 5\pi/8)$ and $\theta\in(11\pi/8, 13\pi/8)$, so that it can wear away only in these two segments. It is also assumed that no undercutting occur in the notch. The actual initial electrode surface are $r=10$ for cathode and $s(\theta,0)=9.5$ for the anode. We have selected $\delta u=0.025$, $\delta\theta=\pi/40$ and $\delta t=0.04$. Figure 1. shows a typical result when the anode is ploted after every ten time steps. Because of symmetry, only the first quadrant is given and in order to magnify the erosion of the anode, the radius r is scaled logarithmicaly according to $r_p = \ln(r/7)$ where r_p is the scaled value of r. The actual depth of the final notch for $s(\pi/2, 2)$ is found to be 7.81 where the corresponding actual depth of final notch obtained by Meyer [7] by using the method of lines is 7.88.

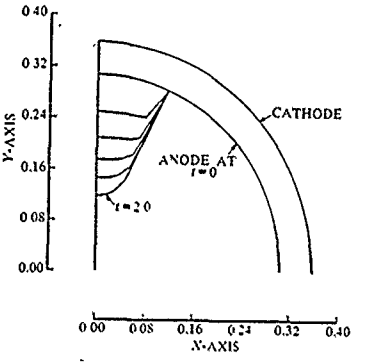


Fig.1. The plot of the anode and cathode surfaces during electrochemical erosion

5. REFERENCES

(1) Boadway, J.D., Int. J. for Numer. Methods in Engng., 10(1976), 527-533

(2) Christiansen, S. and Rasmussen, H., J. Inst. Maths. Applics., 18, (1976), 295-307

(3) Crowley, A.B., J. Inst. Maths. Applics., 24 (1979) 43-57

(4) Elliott, C.M., J. Inst. Maths. Applics., 25(1980), 121-131

(5) Hansen, E.B. and Holm, A.M., Zeitscrift f. angew. Math. U. Mech. 60(1980) 249-251

(6) Hougaard, P., Some solutions of a free boundary problem related to electrochemical machining, Ph.D. thesis, Technical Univ of Denmark (1977)

(7) Meyer, G.H., Numerishe Math, 29, (1978) 329-344

THE DEVELOPMENT OF A COMPUTER ASSISTED LEARNING ENVIRONMENT

Moira J McAlister
Sunderland Business School
1-4 Thornhill Park
Sunderland Polytechnic
Sunderland
Tyne and Wear
England

Dr P Smith
School Of Computer Studies and Mathematics
Sunderland Polytechnic
Green Terrace
Sunderland
SR1 3SD
England

ABSTRACT - This paper presents an integrated, and novel, structure for Intelligent Computer Assisted Learning (ICAL) Systems/Environments.

The environment is a multi-language software platform upon which an interactive CAL system is combined with a set of knowledge/databased systems for student modelling, responding to user's queries, monitoring/diagnosing and assessing the user and the environment. The environment concentrates on the area of Quantitative Methods; and in particular the subject of Linear programming.

Testing has taken place on a formal basis on a wide range of groups from students (e.g. at all applicable level's) to specialists (e.g. cognitive psychologists). These results indicated that 90% of the users favoured the construction of this type of environment. This in itself has been considered a valuable point for the design of a structural methodology for such ICAL environments.

INTRODUCTION - Teaching environments encompass a range of disciplines from Artificial Intelligence, Cognitive Psychology and Linguistics. The breadth of application and the complexity has not always incorporated all areas to their full potential, consequently developments have concentrated upon specific design issues. Examples include SCHOLAR and SOPHIE for modelling issues, WUSOR and QUADRATIC for pedagogical issues, BUGGY and PROUST for diagnostic issues, and NEOMYCIN and ACT for expert systems/cognitive issues.

It is the latter category which has provided the most advanced features and illustrated the effective combination of fields. However, such systems were constructed in modular form and reached very little commercial success; rather sophistication of the domain or specifity of the design placed tremendous restrictions on them. Environments have been created such as Smithtown and Bit-Sized Tutor which attempt to construct an enclosed shell around the user by which subject areas within a curricula are studied.

The lack of acceptance has not been solely a result of the research orientated developments but also the finance and manpower available within the establishments (educational and industrial) for which they were intended. Consequently, this project discusses the construction of a shell, which emulates the best features of a human tutor, whilst redressing the obstacles outlined above.

THE ENVIRONMENT - The environment is named Intelligent Linear Programming Tutoring System (ILPTS) [1] and consists of 4 major sections which were considered to be contained within a human tutor. These were modelling, core teaching material, a query system and a monitoring/diagnostic system. We will consider each of these sections in detail, in each case describing the concepts behind their design.

(1) Modelling - The ability of any system to interact or intercede in a constructive manner which is both useful and helpful to the user is a prerequisite of any system, and was, as stated earlier, a major focus of attention in ILPTS. There are three sections which model a student/user's behaviour. These are:

(a) Bandwidth
(b) Target knowledge
(c) Difference between expert and student/user.

Of the many possible ways of implementing the above 20 have been attempted. These have consisted of the simplest combinations which are the easiest to implement, but do not always exert constructive control over the student/user leading to results which are difficult to assess. ILPTS involves the use of the mental states of the student/user using a declarative target knowledge searching strategy working on a bug-part library with an expert system as a diagnostic program (see Figure 1).

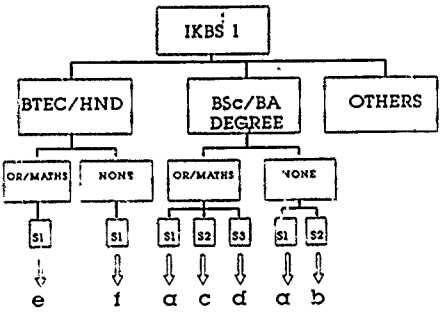


Figure 1

The modelling system, together with the pedagogical interaction, controls the overall operation of the environment and operates via a natural language interface with the user, sitting like a driver on top of the environment. The declarative nature of the modelling system allows a psychological picture of the student/user to be developed,

the incorporation of probability and uncertainty permitting the construction of individualised study programs to be devised. The profile of the student/user is retained and incorporated into the decision process of further study programs when the environment is next used. The nature of such a modelling system permits the detection of misunderstandings and misconceptions which the student/user is perceived to include in their interpretation of the teaching material.

(2) Query - The ability to query the subject area in response to material taught is innate in any lecture/tutorial/seminar session. This not only improves understanding but can also be used to highlight areas of weakness in the light of current material being studied. The system comprises two sections; terminology definition and a problem-solver. These are discussed below.

(a) Terminology Definition - The terminology used within Linear Programming (or, indeed, the majority of areas chosen for ICAL systems/environments) is wide ranging and diverse. The field relates to other areas such as business and computing and is included in light of current curricula. Consequently, the teaching aspects of the environment are enhanced by a system that allows the student/user to obtain the required explanation.

The system focuses on the student/user's input via a natural language interface and is parsed in a manner which allows the system to handle the use of keywords, grammar, context and a combination of keywords.

(b) Problem-Solver - The teaching aspects of many subject areas within ICAL systems/environments involves the use of computers. Linear programming is no exception, and a number of packages are incorporated into a wide variety of curricula. These packages (from mainframe to microcomputer) were studied and the best features of each chosen. One major point which was noted throughout the study was the lack of an appreciation for the dual representation in linear programming.

Consequently, a linear programming problem-solving package was designed and incorporated into the environment. The package includes the following features:-

(1) Development and manipulation of a linear program
(2) Extensive editing facilities
(3) Report generation

All of the above are equally applicable to the primal and dual representations.

This query system compliments the corresponding core CAL material by permitting communication to be a complete two-way process.

(3) Core CAL Material - The pedagogical aspect of the environment is in three sections; text, tutorial and test material which at present is set at two abstraction levels (provisional and professional). The topics covered range from formulation to duality and are in line with the current curricula requirements for which the environment is intended.

(a) Text - The structure of text is vital to the manner in which the student/user will digest the major points to be conveyed.

The structure and consistency of the layout, plus the control of the system ensure that the user absorbs the required material before proceeding.

(b) Tutorial - The same design considerations as with respect to the text material are incorporated but, in this instance the text provided is done so as a sequence of steps which gradually build upon one another giving the student/user a greater appreciation of the corresponding text for that topic and abstraction level.

(c) Test - This subsection interactively interrogates the student/user's understanding of the corresponding topic. The results of the tests, which comprise of questions asked at random, are used to judge the overall effectiveness in the change of the student/user's profile.

(4) Monitoring - This system diagnoses the student/user's understanding. This process occurs by constructing records which indicate the student/user's current profile. Report generation facilities are also available by which individuals/groups (or the environments teaching techniques can be assessed).

IMPLEMENTATION - The environment encompasses a number of features which each have specific requirements. This demanded an effective combination of languages which would allow for the production of an efficient shell by which the concept could be illustrated.

RESULTS - Formal testing was carried out on a wide variety of students/users from a range of backgrounds which encompassed the skills necessary to assess such an environment.

CONCLUSIONS - The overall aim of the project was to create an environment which best emulated an ideal human tutor, and as such placed emphasis on modelling, pedagogical, and diagnostic processes as and when necessary. The hope is that such a development will assist in leading to the construction of a design methodology for such systems, while allowing ideas within the environment to be advanced still under the belief that a combination of relevant fields are necessary for such ideas to illustrate practical progress.

REFERENCES

1. An Integrated Structure For ICAL Systems, M J McAlister and Dr P Smith, submitted to Journal Of Computer Assisted Learning, 1991.

A LINEAR SYSTOLIC ARRAY FOR THE QUOTIENT-DIFFERENCE ALGORITHM

Octav Brudaru
Universitatea "Al. I. Cuza" Iasi
Seminarul Matematic "A. Myller"
6600-Iasi, Romania

Abstract. It is presented a linear systolic array for the column-by-column version of the quotient-difference algorithm for the polynomial roots finding. It is proved the correctness of the design and it is computed the bus bandwidth and processor utilization. The asymptotic performances of the array are also estimated and optimized.

1. Introduction

In this paper it is presented a linear systolic array implementing the column-by-column version of the quotient-difference (CCQD) algorithm for the solution of the equation

$$p(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n=0, \qquad (1)$$

where p is a polynomial with real coefficients. If $(x_k)_{k\geq -n+1}$ is the solution of the associated difference equation $a_0x_k+a_1x_{k-1}+\ldots+a_nx_{k-n}$, for which $x_j=0$, $j=n+1,\ldots,-1$ and $x_0=1$, then the CCQD algorithm involves the computing of the following recursions ([5,pp.321]) :

$$q(k,j+1)=d(k+1,j)/d(k,j)q(k+1,j), \qquad (2)$$

$$d(k,j)=q(k+1,j)-q(k,j)+d(k+1,j-1), \qquad (3)$$

$j=1,2,\ldots,n-1$, where $q(k,1)=x_{k+1}/x_k$ and $d(k,0)=0$, for $k=0,1,\ldots,K$.

If no two consecutive roots of p have the same absolute value, then $\lim q(k,j)=r_j$, for k tending to infinity and $p(r_j)=0$, $j=1,\ldots,n$, $r_j<r_{j+1}$, $j=1,\ldots,n-1$.

In [4] it is presented a systolic design implementing the row-by-row version of the quotient-difference algorithm.

In section 2 we present a linear systolic array implementing (2) and (3) and give a correctness proof of its working. In section 3 we analyse the performances of the array, while some variants of the basic design are discussed in the last section.

2. A linear systolic array

We define the clock tick (CT) as the time to execute a division and suppose that each processing element (PE) is active during every CT. The time is denoted by t and represents the number of CTs. Also, L(t) denotes the value circulating through the pin having the label L during the t-th CT.

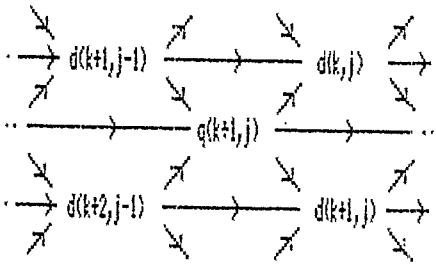


Figure 1: The precedence constraints

The digraph representing the precedence constraints in the computation defined by (2) and (3) is represented in Figure 1.

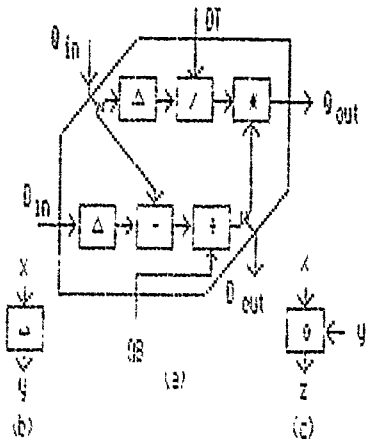


Figure 2: The basic systolic cell

The structure of the basic systolic cell (BSC) that one uses in the sequel is given in Figure 2(a). The PEs in BSC are depicted in Figure 2(b)-(c). The PE in Figure 2(a) performs $y(t+1)=x(t)$ while, the PE in Figure 2(b) executes $z(t+1)=x(t)oy(t)$, where $o\in\{+,-,*,/\}$. The working of BSC is stated by

Proposition 1.

$$D_{out}(t+2)=D_{in}(t-1)-Q_{in}(t)+QB(t+1) \quad (4)$$

and

$$Q_{out}(t+3)=Q_{in}(t)/DT(t+1)D_{out}(t+2) \quad (5)$$

for any $t>0$.

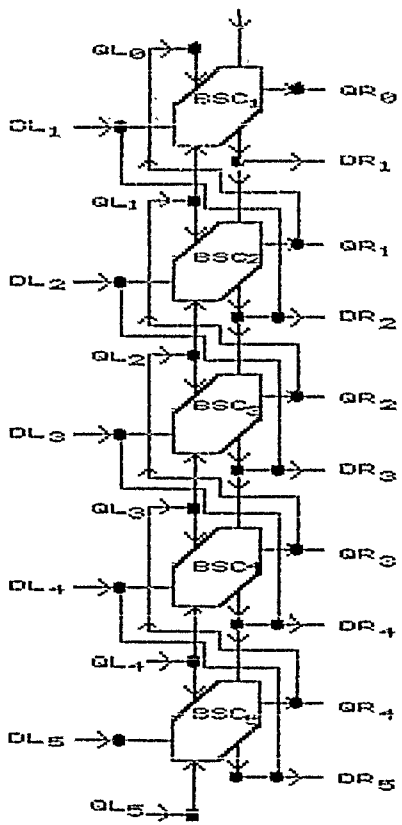


Figure 3: The linear systolic array

The linear systolic array (LSA) implementing the CCQR method is obtained by serially connecting K BSCs, BSC_k, $k=1,\dots,K$, as it is shown (for $K=6$) in Figure 3. The input data is introduced through $QL_0,\dots,QL_K$ (the initial q-values) and $DL_1,\dots,DL_K$ (the initial d-values). The computed q-values are extracted from $QR_0,\dots,QR_{K-1}$. These values are sent back to the left entries in order to advance in the computation with a new column both for d and q. Let us remark that CCQR method requires to compute $d(k,i)$, $k=0,1,\dots,K-2i$, and $q(k,i+1)$, $k=0,1,\dots,K-2i-1$, where $i=1,\dots,n-1$. Since we need at least one term of the sequence $(q(k,n))_{k\geq 0}$, it obtain that $K\geq 2n-1$.

Further, one denotes by $L(k;t)$ the value circulating through the pin L in BSC_k at the time t. The correctness of the working of LSA is stated by

Theorem 2. If

$$QL_k(t+k)=q(k,1),\ k=0,\dots,K \quad (6)$$

and

$$DL_{k+1}(t+K-1)=d(k+1,0),\ k=0,\dots,K-1 \quad (7)$$

then

$$QR_{k+1}(t+k+4i)=q(k,i+1),k=0,1,\dots,K-2i-1 \quad (8)$$

and

$$DR_{k+1}(t+K-2+4i)=d(k,i),\ k=0,1,\dots,K-2i \quad (9)$$

for $i=1,\dots,n-1$.

Proof. The proof is done by induction. For the beginning we prove (8) and (9) for $i=1$. Let us examine the working of BSC_1. From (6) and (7) we have $Q_{in}(1;1)=QL_0(1)=q(0,1)$, $D_{in}(1;t-1)=DL_1(t-1)=d(1,0)$, $QB(1;t+1)=QL_1(t+1)=q(1,1)$. From (4) it results $DR_1(t+2)=Q_{out}(1;t+2)=Q_{in}(1;t-1)-Q_{in}(1;t)+QB(1;t+1)=d(1,0)-q(0,1)+q(1,1)$, and with (3) we obtain $DR_1(t+2)=d(0,1)$. Now, suppose that $DR_{k+1}(t+K+2)=d(k,1)$, for $k\geq 0$ and examine the activity of BSC_{k+2}. From (6) and (7) we obtain $Q_{in}(k+2;t+k+1)=QR_{k+1}(t+k+1)=q(k+1,1)$, $D_{in}(k+2;t+k)=QL_{k+2}(t+k)=d(k+2,0)$, $QB(k+2;t+k+2)=QL_{k+2}(t+k+2)=q(k+2,1)$ and from the above assumption $DR(k+2;t+k+2)=Q_{out}(k+1;t+k+2)=QR_{k+1}(t+k+2)=d(k,1)$. By taking $t+k+1$ instead of t in (4), it results $D_{out}(k+2;t+k+3)-D_{in}(k+2;t+k)-Q_{in}(k+2;t+k+1)+QB(k+2;t+k+2)=d(k+2,0)-q(k+1,1)+q(k+2,1)$ and from (3) we obtain $DR_{k+2}(t+k+3)=d(k+1,1)$. Consequently (9) holds for $i=1$.

On the other hand, $QR_{k+1}(t+k+4)=$

$Q_{out}(k+2;t+k+4)$ and from (5) it results $QR_{k+1}^{out}(t+k+4)=Q_{in}(k+2;t+k+1)/DT(k+2;t+k+2)=D_{in}(k+2;t+k+3)=(q(k+1,1)/d(k,1))d(k+1,1)=q(k,2)$ i.e. (8) holds for i=1.

Now, assume that (8) and (9) are true for some i. From the construction of LSA, we have $QR_{k+1}=QL_k$, and $DR_{k+1}=DL_k$, k=0,...,K-1. Consequently, $QL_k(t+k+4i)=QR_{k+1}(t+k+4i)=q(k,i+1)$, k=0,...,K-2i-1, and $DL_k(t+k+4i-2)=DR_{k+1}(t+k+4i-2)=d(k,i)$, k=0,...,K-2i. BSC$_{k+1}$ receives $Q_{in}(1;t+4i)=QL_0(t+4i)=q(0,i+1)$, $D_{in}(1;t+4i-1)=DL_0(t+4i-1)=d(1,i)$, $QB(1;t+4i+1)=QL_1(t+4i+1)=q(1,i+1)$ and $DT(1;t+4i+1)=0$. So, following (4), we obtain $DR_1(t+4i+2)=D_{out}(1;t+4i+2)=D_{in}(1;t+4i-1)-Q_{in}(1;t+4i)+QB(1;t+4i+1)=d(1,i)-q(0,i+1)+q(1,i+1)=d(0,i+1)$. Now, let us suppose that $DR_k(t+k+4i+1)=d(k-1,i+1)$, $k\geq 1$. BSC$_{k+1}$ receives $Q_{in}(k+1;t+k+4i)=QR_{k+1}(t+k+4i)=q(k,i+1)$, $D_{in}(k+1;t+k+4i-1)=DR_{k+1}(t+k+4i-1)=d(k,i)$, $QB(k+1;t+k+4i+1)=QL_{k+1}(t+k+4i+1)=q(k+1,i+1)$, and $DT(k+1;t+k+4i+1)=DR_k(t+k+4i+1)=d(k-1,i+1)$. So, following (4), $D_{out}(k+1;t+k+4i+2)=D_{in}(k+1;t+k+4i-1)-Q_{in}(k+1;t+k+4i)+QB(k+1;t+k+4i+1)=d(k+1,i)-q(k,i+1)+q(k+1,i+1)$. From (3) it results $DR_{k+1}(t+k+4i+2)=d(k,i+1)$. On the other hand, using (5) we obtain $QR_k(t+k+4i+3)=Q_{out}(k+1;t+k+4i+3)=Q_{in}(k+1;t+k+4i)/DT(k+1;t+k+4i+1)D_{out}(k+1;t+k+4i+2)=q(k,i+1)/d(k-1,i+1)d(k,i+1)$ and from (2) it results $QR_k(t+k+4i+3)=q(k-1,i+2)$, and the proof is terminated. #

3. The performances of the systolic array

Let us suppose that a single polynomial p(x) is to be processed on LSA. The computation starts at tmin=t-1 with the introducing of q(0,1)=0 through QL_0 and ends at tmax=t+K+2n-3 when q(K-2n+1,n-1) is extracted from QR_{K-2n+2}. Thus the computation takes ttot=tmax-tmin+1=K+2n-1 CTs. From (8) it results that a new q-value emerges from each QR-output at every four CTs. This means that each BSC in LSA has a poor utilization.

Now, assume that we have to find the roots of the polynomials p_s, s=1,...,S, where S=4M. One denotes by B_i the i-th batch, where $B_i=(p_j/j=4i-3,\ldots,4i)$, i=1,...,M. For the sake of the regularity we assume that all polynomials are of degree n where $n=\max(n',n'\leq(K+1)/2)$, this choice of n being given by the condition $K\geq 2n-1$ (the length K of LSA is supposed to be fixed).

The processing of B_i, i=1,...,M is done in a serial fashion. For each fixed i, the processing of B_i begins with the introducing of the q-values of p_{4i-3} in the manner indicated by Theorem 1, j=4i-3,...,4i, so that the q(0,1)-value of p_{j+1} is introduced one CT after the introducing of q(0,1)-value of p_j. One passes to B_{i+1} as soon as the q(o,n)-value of p_{4i} emerges from QR_1. Further we shall estimate the effectiveness with which the resources of LSA are used when it processes a single batch of four polynomials. Let θ denote the time spent for arithmetic operations by a single processor algorithm (SPA) which acts on a single problem instance. From (8) and (9) it results that SPA computes (n-1)(K-n) values of q and (n-1)(K+1-n) values of d. Therefore, the total number of arithmetic operations is Q=2(n-1)(2k-2n+1), and consequently SPA needs T(SPA)=4xQ time for each batch. On the other hand, LSA takes T(LSA)=ttot+3=k+2n+2 CTs per batch, and has Np=4K processors. The ratio Ec=T(LSA)xNp/T(SPA) measures the effectiveness of processor utilization ([2]) and we obtain

$$Ec(K,n)=K(K+2n+2)/(2(n-1)(2K-2n+1)). \quad (10)$$

Let us suppose that K=an+b, where a and b are integers. Further, we are interested to obtain a_0 and b_0 so that $\lim_{n\to\infty} Ec(a_0n+b_0,n)\leq\lim_{n\to\infty} Ec(an+b,n)$, where a_0n+b_0, $an+b\leq 2n-1$, for each $n\geq 3$. From (10), we obtain $\lim_{n\to\infty} E_c(an+b,n)=f(a)=a(a+2)/(4(a-1))$ while a and b satisfy either (i) $a\neq 2$ and $b\geq -1$ or (ii) $a>2$ and $(b+1)/(2-a)\leq 3$. Finally, it results $f(3)<f(a)$, for $a\geq 2$ and $a\neq 3$, and we obtain $a_x=3$, $b_x\geq -4$ and $f(3)=1.875$.

Now, let us estimate the bus bandwidth utilization, Ed=WxNd/Ad ([2]). The maximum amount of data to be transmitted between the host and LSA in one CT is W=5. The data movement begins with q(0,1) of the first polynomial at time t and ends with q(K-2n+1,n) of the last polynomial of the batch at the time t+K+2n. Thus, the number of data

movement steps is Nd=K+2n+1. The processing of a polynomial needs K+2 input data and offers n+1 final q-values. Thus, the total amount of data to be transmitted is Ad=4(K+n+1). Consequently, Ed(K,n)=5(K+2n+1)/[4(K+n+1)] and $\lim_{n\to\infty} Ed(3n+b,n)=25/16$. The data locality ratio ([3]) is RDL=Nmem/Nop, where Nmem is the number of memory acceses and Nop the number of floating point operations. It is to see that Nmem=Nd, Nop=T(SPA) and for any $K\geq 2n-1$ we have $\lim_{n\to\infty} RDL(K,n)=0$, while for each fixed n it results $\lim_{K\to\infty} RDL(K,n)=1/[16(n+1)]$.

4. Some variants

The CCQD algorithm requires to supply the array with the initial q-values. These values can be obtained from the systolic array, implementing the Bernoulli's method for polynomial roots finding ([1]).

It is clear that the array could be reconfigurated in order to obtain a trapezoidal systolic array. In this case, LSA is modified so that the connections for recirculating the outputs are removed. The 2D systolic array is obtained by using I copies of LSA, $LSA_1,\ldots,LSA_I$, so that LSA_i (wich has K-i+1 BSCs) sends the results to LSA_{i+1}, i=1,...,I-1, while LSA_I emits back to LSA_1. We remark that for sufficiently large I the data recirculating can be removed.

By connecting BSC_K to BSC_1 it is obtained a ring of processors wich is able to handle with larger values of K. A torus can be obtained by applying this idea to 2D array above suggested.

A new variant of LSA could be obtained by supposing that the initial q- and d-values enter LSA at the same time. The network must be retimed and many delay processors can be saved.

A more flexible variant of LSA (supporting any manner to introduce the initial data) can be obtained by using the so-called diastolic control elements ([6]). It suffices to replace each delay register in the circuit of BSC by a diastolic control element. In this manner we ensure a local control of LSA wich becomes a systaltic array ([6]).

References

[1] Brudaru, O., Systolic Arrays to Solve Algebraic Equations by Bernoulli's Method, Publications de l'Institut Mathematique, Nouvelle serie, 44(58),1988,pp.137-142.

[2] Cheng, K. H., VLSI Systems for Band Matrix Multiplication, Parallel Computing, no. 4,1987, pp.239-258.

[3] Chronopoulos, A.T., Gear, C.W., On the Efficient Implementation of Preconditioning s-Step Conjugate Gradient Methods on Multiprocessor with Memory Hierarchy, Parallel Computing no. 11, 1989, pp. 37-53.

[4] Evans, D.J., Megson, G.M., Systolic Array for the Quotient Difference Algorithm, IEE Proceedings, vol. 135, Pt. E, no. 1, January 1988.

[5] Scheid, D.P., Theory and Problems of Numerical Analysis, Shaum's Outline Series, McGraw-Hill, New York, 1968.

[6] O'Leary, D.P., Stewart, G.W., From Determinancy to Systaltic Arrays, IEEE Transactions on Computers, vol. C-36, no. 11, November 1987.

Acknowledgement

Many thanks to Prof. Dr. Calin Ignat for his constant encouragement of this work and for his valuable suggestions in the preparation of the paper.

BONDGRAPH MODELLING AND SIMULATION OF A CLOSED LOOP POWER AMPLIFIER

S.R.Bhat & L.Umanand
Centre for Electronics Design & Technology (CEDT)
Indian Institute of Science, Bangalore-560 012, INDIA.

Abstract: In this paper the modelling of a closed loop power amplifier is described using the Bondgraph techniques. The power amplifier is a 50 watt class AB pushpull linear amplifier, which supplies power at 230v ac rms. It draws power from a 12v battery. The bondgraph modelling and simulation is performed on a personal computer using the "TUTSIM" software package. The power amplifier is also realised in the laboratory and the results are compared with the simulated model.

1. Introduction

Bondgraph is a graphical tool which is used in the representation of dynamic systems. It displays both the energy and signal exchanges between components or elements of the system [1]. For a system of order greater than three, the transfer function approach to analysis becomes cumbersome and also one tends to loose the correspondence of the various time constants associated with the various elements in the system, especially for high order systems. But in the case of bondgraphs, one to one correspondence is maintained with the actual circuit, and further, it gives a graphical picture of the dynamic equations of the system. The bondgraphs are initially set up acausally and then subsequently tranformed into a causal diagram by a systematic choice of the causalities [1,2].

2. System and Model

The circuit shown in fig.1 is the system which has been modelled and simulated. It is a linear pushpull amplifier operated in the class AB mode. The drive for the transistors is from a push pull stage driven from an integrated amplifier. A sinewave reference of frequency 50Hz, drives the bases of the pushpull transistors through a transformer TX1 (fig.1). The power stage consists of two power transistors and a step up transformer TX2. The output voltage fed to the load is 230 v rms at 50 Hz. The output voltage is maintained at 230 v by means of feedback loop as shown in fig.1.

The transformer TX2 is modelled taking into account the leakage and magnetising inductances, and winding resistances.The transistor is represented by a large signal equivalent circuit, as it handles power signals. It is modelled as a voltage controlled current source through the Ic-Vbe non-linear characteristics. The diode is represented by its nonlinear i-v characteristics.

The TUTSIM package [3] is used in this simulation on a personal computer. As there are no representations for transistor and diode in the bondgraph language, the model for transistor and the diode (as mentioned above) were built using the various "Function blocks" provided in TUTSIM. The transistor is represented as a dependent current source (Qf) as shown in fig.2(a) and the diode is represented as a piecewise linear resistor (D) as shown in fig.2(b). The bondgraph model of the entire system is shown in fig.3.

3. Simulation Results

One of the difficulties faced during design of a closed loop control system is the tuning of the controller parameters for optimum control performance. Bondgraph simulation is a very useful tool which helps one to tune the controller and the sensor time constants to achieve optimum control performance, as it gives directly the required signals in time domain and the parameters can be changed quickly during the simulation.

The simulation results are shown in figures 4(a) and 4(b). The fig.4(a) shows the output voltage waveform (Vo) and the sense filter output waveform (Vc) during the transient period and fig.4(b) these signals during the steady state period. Simulation results indicate that for stable operation, the sense time constant must be much faster than the controller time constant. As the sense time constant approaches the controller time constant, the system becomes increasingly unstable. It was found from this simulation that a sense time constant which is twice as fast as the controller time constant gives a stable operation.

The power amplifier circuit was implemented in the laboratory to verify the above simulation results. The simulation proved to be very accurate. Through simulation, a sense time constant of 273ms and a controller time constant of 384ms provided a very stable system from no load to full load and was used in the hardware implementation successfully. To verify the simulated results further, the value of the sense time constant in the hardware unit was varied around the optimum value. The system proved to be unstable as was predicted during simulation. Further, the observed output waveforms and the simulated output waveform agreed in shape and amplitude.

4. Conclusion

In the above discussion, we looked at modelling and simulating a pushpull amplifier system, with the objective of analysing the system and to tune the controller time constants to achieve optimum control performance. This kind of simulation reduces the hardware implementation time considerably as one can simulate the non-linear devices like transistors and diodes, one can predict the nature of the outputs, design the control circuit to meet the required performance, etc. in a short time.

References

1. Introduction to Bondgraphs & their applications by Dr.Thoma, Pergamon press, 1975.
2. System Dynamics: A unified approach, by Karnopp & Rosenberg, John Wiley & Sons, 1975.
3. TUTSIM manual version 6.55

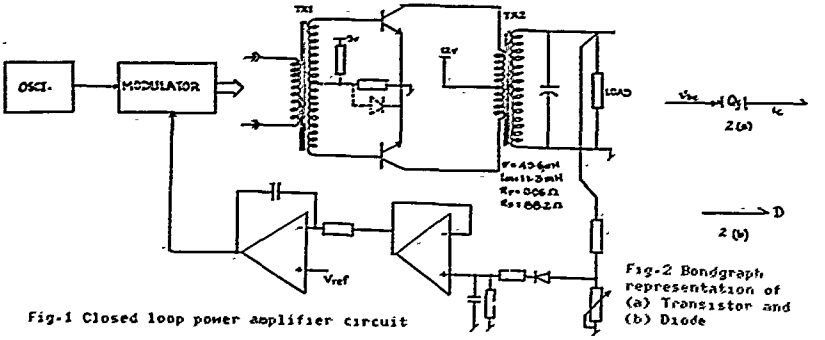


Fig-1 Closed loop power amplifier circuit

Fig-2 Bondgraph representation of (a) Transistor and (b) Diode

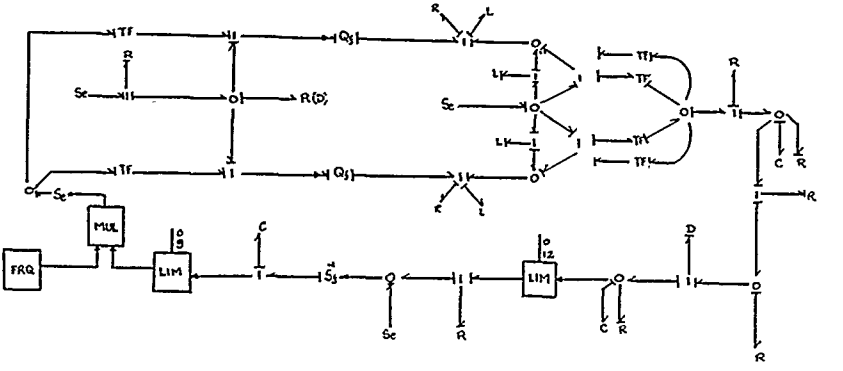


Fig-3 Simplified Bondgraph Model

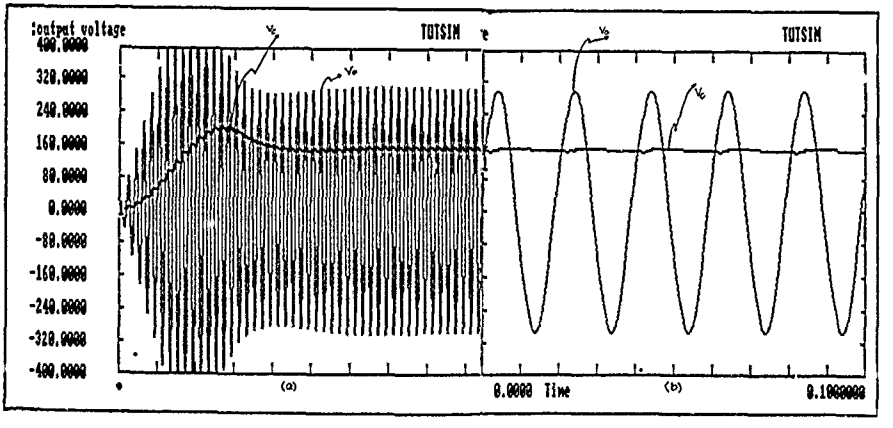


Fig-4 Simulation results (a) Transient (b) Steady state

ERROR ESTIMATION AND H-REFINEMENT ADAPTIVE TECHNIQUES FOR ELECTROMAGNETIC ANALYSIS APPLICATION

P. FERNANDES(*), P. GIRDINIO(+), G. MOLINARI(+),

(*) Istituto per la Matematica Applicata del CNR, Via L.B. Alberti 4, 16132 Genova, Italy
(+) Dipartimento di Ingegneria Elettrica - Università di Genova, Via Opera Pia 11a, 16145 Genova, Italy

Abstract This work presents a comparative survey of h-refinement adaptive schemes for the solution, by means of a first order finite element method, of the differential equations of interest in computational electromagnetics. All schemes analysed are based on "a posteriori", element by element, single solution error estimation techniques. A number of possible algorithms in the above class is reviewed, and obtained results are compared and discussed.

INTRODUCTION

The quality of solution that can be obtained by finite element methods is strongly affected by the mesh used. For this reason, the definition of a mesh suitable for a particular problem has always been, and in most practical cases still is, one of the most crucial phases of a finite element analysis. As a consequence, a very significant research effort is devoted at present to develop suitable adaptive meshing algorithms, that are recognized of critical importance to relieve the analyst from this difficult and time consuming task and to make quality of solution independent of user skillfullness in meshing.

The first attempts in this direction have been performed in the area of structural analysis, historically the first to develop finite element techniques, where the most commonly used elements are quadrilateral ones, for this reason, many of the most significant papers in the literature about adaptive mesh ing deal with quadrilateral elements [1] However, the aim of the authors is to develop adaptive meshing algorithms and procedures suitable for application in electromagnetic analysis, or computational electromagnetics, that has a number of different requirements. In fact, on the contrary of structural analysis, computational electromagnetics must frequently solve open boundary problems, has a significant variety of differen tial equations to face, frequently nonlinear and time-dependent, and in application of practical interest must cope in general with intricate geometries and a number of different materials For this class of applications triangular meshes are generally used, mainly because of the greater simplicity in dealing with intricate boundaries and interfaces. However, this kind of elements has received so far much less attention in adaptive mesh ing literature [2].

Within computational electromagnetics, the attention of the authors has been focused on adaptive meshing algorithms suitable for "general-purpose" analysis codes, that is, codes capable to solve a wide class of geometries of practical interest, generally unknown "a priori"to the code developer, and then to the meshing algorithm Because of this further specific requirement, a number of stringent features are necessary or desirable In particular, the adaptive meshing algorithm should be gen eral enough to be suitable, or readily adapted, to a number of differential equations, also nonlinear and time-dependent, and should be "robust" enough to perform correctly even with se vere geometrical constraints. Finally, since this class of applications is significantly compute-bound, computationally intensive algorithms should be as far as possible avoided. As a consequence, no activity has been performed on adaptive methods based on dual finite element solutions [6], judged not general enough and too computationally expensive, nor on methods requiring the solution of a local differential problem on "finite patches" made up with several elements [7], since they involve an "ad hoc" treatment of boundaries and interfaces, that can become cumbersome or even critical in intricate geometries. In order to maximize flexibility and robustness, the activity of the authors on error estimation and adaptive meshing has then been concentrated on first order triangular elements, using element-by-element error estimation and h-refinement techniques [3-5].

In the present paper a review of the error estimation and adaptive meshing algorithms and techniques selected and implemented is performed, and results of some comparative analysis of their features are presented and discussed.

MESH REFINEMENT ALGORITHM

An h-refinement adaptive meshing algorithm is organized as follows. first a finite element solution is computed on the current mesh, then an estimation of the local error of this solution is computed element by element, finally the elements showing greatest errors are subdivided to obtain a new mesh. Since the various methods in this paper differ only for the error estimation procedure used, in the present section mesh refinement algorithm common to all of them is described.

To start the adaption procedure, an initial mesh is built up using the automatic meshing features of the CEDEF package [8], a first problem solution is performed, the local error on each element is estimated and used to decide which elements are to be subdivided. Once an element has been marked for refinement, two different element subdivision rules have been used, for elements lying in the bulk or having a side on a boundary or on an interface, respectively.

In the first case, a node in the centroid of the element is added, splitting the original element in three smaller ones. For boundary and interface elements, a node is added at the midpoint of the boundary or interface side. In this way it is possible to thicken the mesh also on boundaries and interfaces, which would not be affected otherwise using the first technique only. Once this procedure has been applied to the whole mesh, a Delaunay triangulation is performed, to provide the optimal mesh for the given set of nodes, and a new solution is computed. The procedure is then iterated until the convergence criterion is satisfied.

The final step of mesh improvement is an iterative nodal displacement procedure, operated on patches of elements with an underrelaxed "rubber banding" technique, that can be activated by the user to improve the final quality of the mesh before performing the final solution.

ERROR ESTIMATION PROCEDURES

The three error estimator procedures, which in the previous work of the authors have emerged as the best ones to be used in the adaptive mesh algorithm, are described in the present section. Two of them are based on the residual evaluation and are referred to as "local error problem" and "extended complete residual" methods [3,5], while the other one is based on polynomial interpolation theory and is referred to as "field difference" method [3,4]. The "local error problem" is as follows. Considering a Poisson problem defined by the differential equation:

$$-\nabla^2 u = f \tag{1}$$

in a region Ω with boundary conditions on $\partial\Omega = \Gamma = \Gamma_n \cup \Gamma_d$:

$$u = u_d \quad on\ \Gamma_d \qquad \frac{\partial u}{\partial n} = q \quad on\ \Gamma_n \tag{2}$$

the local error e_i on the i-th element Ω_i is estimated by solving the following differential problem on Ω_i:

$$-\nabla^2 e_i = r \qquad on\ \Omega_i \tag{3}$$

$$e_i = r_d \quad on\ \partial\Omega_i \cap \Gamma_d \qquad \frac{\partial e_i}{\partial n} = r_n \quad on\ \partial\Omega_i \cap \Gamma_n \tag{4}$$

$$\frac{\partial e_i}{\partial n} = \frac{1}{2}\left[\left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{out} - \left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{in}\right]_{\partial\Omega_i - \Gamma} \quad on\ \partial\Omega_i - \Gamma \tag{5}$$

where

$$e_i = u - \tilde{u} \qquad r = f + \nabla^2 \tilde{u} \tag{6}$$

$$r_d = u_d - \tilde{u} \qquad r_n = q - \frac{\partial \tilde{u}}{\partial n} \tag{7}$$

and $\tilde{u}$ is the finite element solution of the problem (1) - (2). Once the error e_i on the solution has been estimated, the error on its gradient (i e. $\nabla u - \nabla\tilde{u}$) is readily obtained as ∇e_i. Then the refinement indicator η_i on the element Ω_i is computed as [4]:

$$\eta_i = \left(\kappa \frac{\|\nabla e_i\|^2}{\sum_{j=1}^{N} \|\nabla \tilde{u}_j\|^2} + (1-\kappa)\frac{\|e_i\|^2}{\sum_{j=1}^{N} \|\tilde{u}_j\|^2}\right)^{\frac{1}{2}} \tag{8}$$

where N is the total number of elements in the whole domain Ω and κ is a weighting factor to be selected in the range 0 to 1, allowing to define it as the error estimator on the solution, on its gradient, or on both. The values $\eta_1, \ldots, \eta_N$ are then used to decide which elements will be subdivided. The quadratic norm in (8) is the standard one, that is to say:

$$\|v_i\| = \sqrt{\int_{\Omega_i} |v_i|^2 dS} \tag{9}$$

where v_i is a generic scalar or vector quantity in Ω_i. The refinement criterion used is based on the following procedure. the maximum value of the refinement indicator over the whole domain, η_{max}, is computed after each problem solution, and an element is marked for subdivision at the next refinement iteration when:

$$\eta_i > \sigma \eta_{max} \tag{10}$$

where σ is a user defined refinement parameter such that $0 < \sigma < 1$.

The "extended complete residual" method assumes that an estimate of the error on the solution, s_i, can be obtained by combining the driving functions of the "local error problem" as:

$$s_i = 2\gamma\left(\int_{\Omega_i} |r|\, dS\right) + 2(1-\gamma)\left(\int_{\partial\Omega_i - \Gamma} \frac{1}{2}\left|\left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{out} - \left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{in}\right| dl + \int_{\partial\Omega_i \cap \Gamma_n} |r_n| dl\right) \tag{11}$$

where the weight γ ($0 \le \gamma \le 1$) allows to weight differently the error in the element and on its boundary. The estimator s_i has the meaning of an average error over Ω_i, but to use the same refinement indicator as in the "local error problem" method we need rather an estimate of the square of the quadratic norm of the error, that is evaluated as:

$$\|e_i\|^2 = |s_i|^2 A_i \tag{12}$$

where A_i is the area of the element Ω_i.

To provide also an estimate of an average error on the gradient of the solution on the i-th element, the quantity $\vec{g}_i$ has been defined as:

$$\vec{g}_i = \vec{a}_x g_{ix} + \vec{a}_y g_{iy} \tag{13}$$

where g_{ix} and g_{iy}, for an element without any side on the boudary, are obtained by a least square solution of the following overdetermined linear system:

$$(\vec{a}_x \cdot \vec{n}_{ij}) g_{ix} + (\vec{a}_y \cdot \vec{n}_{ij}) g_{iy} = \frac{1}{2}\left[\left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{out} - \left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{in}\right] \qquad (j = 1,2,3) \tag{14}$$

If any side of Ω_i is on a Neumann boundary the corresponding equation in the system (14) is:

$$(\vec{a}_x \cdot \vec{n}_{ij}) g_{ix} + (\vec{a}_y \cdot \vec{n}_{ij}) g_{iy} = r_n \tag{15}$$

If one side of Ω_i is on a Dirichelet boundary, the system (14) has two equations only, so that it is no longer overdetermined and an ordinary solution exists. If two sides of Ω_i are on a Dirichelet boundary, $\vec{g}_i$ is directly determined by assuming that it is directed as the $\vec{n}_{ij}$ normal to the only internal side, that is to say:

$$\vec{g}_i = \frac{1}{2}\left[\left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{out} - \left.\frac{\partial \tilde{u}}{\partial n_{ij}}\right|_{in}\right]\vec{n}_{ij} \tag{16}$$

From $\vec{g}_i$ the square of the quadratic norm of the error on the gradient of the solution, needed in (8), is readily estimated as:

$$\|\nabla e_i\|^2 = |\vec{g}_i|^2 A_i \tag{17}$$

For the "field difference" method approach, let us first define as $\vec{E}_i^{(e)}$ the gradient of the solution in a generic element i, constant over the element, and $\vec{E}_j^{(n)}$ the gradient of the solution assigned to the generic node j, computed as the average of the $\vec{E}_i^{(e)}$ values of the elements surrounding the node and having the same material properties. The error on the gradient of the solution in the node j of the element i has then be defined as.

$$\nabla e_{ji} = \vec{E}_i^{(e)} - \vec{E}_j^{(n)} \tag{18}$$

This quantity can be defined in any node of each element, and can assume different values for the same node when computed for elements lying across an interface between different materi-

ls. By assuming that this error, computed in the nodes only, is distributed over the element with the element shape function N_j, it is possible to define the function $\nabla e_i = \sum_{j=1} N_j \nabla e_{ji}$, giving a distribution of the error over the whole element. The square of the quadratic norm of the estimate of the error on the gradient of the solution over the generic element i is given by:

$$\|\nabla e_i\|^2 = \int_{\Omega_i} |\nabla e_i|^2 dS \qquad (19)$$

Since the "field difference" algorithm defined in this way provides only an estimate of the error on the gradient of the solution, the weighting factor κ of eq. (8) has been set to 1.

At the end of the adaption process, the relative estimated error on the solution and on its gradient, relevant to the element Ω_i, according to the quantities made available by the basic error estimator routines, are computed as:

$$e_{ei} = \frac{\|e_i\|}{\|\tilde{u}_i\|} \qquad (20)$$

$$e_{gi} = \frac{\|\nabla e_i\|}{\|\nabla \tilde{u}_i\|} \qquad (21)$$

TEST STRATEGY AND RESULTS

Preliminary tests have shown that it is difficult to perform significant comparisons among adaptive meshing algorithms even if they differ only in the method of error estimation. In fact, the error estimation method usually determines both the choice of the elements to be refined and the iteration at which the algorithm stops. Since the best method for the selection of the elements to be refined can be in principle different from the best one to be used as convergence criterion of the refinement, the two effects must be separated so as to work out significant comparisons. In order to do so, a test case provided with analytical solution has been used, stopping the mesh refinement when the greatest difference between the analytical and numerical solutions (at the nodes for the potential and at the centroid of the element for the field) falls below a fixed limit. Thus the error estimation methods are used only to choose the elements to be refined. In this way, the comparison of "true errors" and CPU time directly shows the relative merit of each error estimation method used as a refinement indicator. Of course, different methods give solutions on different final meshes, so that comparing the deviations of the estimated error with respect to the "true" one is not fully significant. Thus, evaluating the quality of the method as error estimator, such a comparison has been carried out on the same mesh for all methods.

As a first step to perform significant comparisons among the various methods, each of them has been optimized with respect to the parameters involved, σ and κ. To this aim, the results obtained for different values of the same parameters have been compared and a compromise taking into account both the execution time and the "true" relative error on the potential or on the field has been done to choose the "optimal" parameter value. The "true" relative errors have been defined as:

$$e_{eti} = \frac{\|e_{ti}\|}{\|u_i\|} \qquad (22)$$

$$e_{gti} = \frac{\|\nabla e_{ti}\|}{\|\nabla u_i\|} \qquad (23)$$

where u_i is the analytical solution over the generic element Ω_i, ∇u_i its gradient, $e_{ti} = u_i - \tilde{u}_i$, and $\nabla e_{ti} = \nabla u_i - \nabla \tilde{u}_i$. The error (22) or (23) is used depending on whether the quantity of interest is the potential or the field. The same criterion has been used in comparing the best version of each method, for the selection of the best method as refinement indicator. Finally, the relative error (20) or (21) on the mesh obtained by the best refinement indicator has been evaluated by each method and compared with the "true" relative error (22) or (23). Such a comparison provides information about the relative merit of each method when used as a tool to assess the accuracy of the final results. The whole procedure has been carried out using both the errors (20) and (22) on the potential and the errors (21) and (23) on the field, since the best refinement indicator to compute the potential is not necessarily the best one to compute the field. A similar argument holds when the methods are regarded as estimators of the final error.

The solution of the Laplace equation in an L-shaped region has been used as test case. This region is shown in Fig. 1 with the boundary conditions used and the analytical solution superimposed. The $A - B$ line, along which the errors used to compare the various methods have been computed, is also shown in Fig. 1. All tests have been carried out starting from the initial mesh shown in Fig. 2.

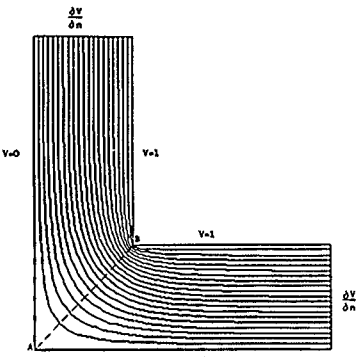


Fig. 1 - Equipotential lines of the analytical solution of the test case.

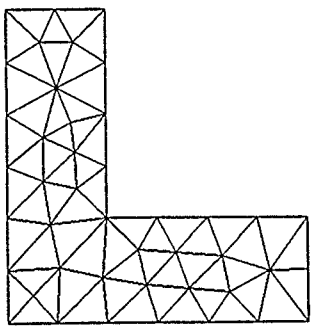


Fig. 2 - Initial mesh used for all tests.

As a first step, the σ parameter in equation (10) has been optimized for each method according to the methodology previously described. Once the best σ has been found, also the κ parameter in equation (8) has been similarly optimized, except for the "field difference" method which only allows $\kappa = 1$. In Fig. 3 the "true" relative errors (22) on the potential, along the $A - B$ line, on the solution obtained by using the various methods as refinement indicators, each with its best value of σ and κ, are shown.

In order to compare the methods as estimators of the final errors, the relative error estimation of eq. (20) on the solution provided by the final mesh of the "extended complete residual" method, which has provided the best results, has been performed for each method along the previously defined line $A - B$ of Fig. 1. The difference between the "true" and the estimated errors (22) and (20), along the $A - B$ line, is shown for each method in Fig. 4.

In Figs. 5 and 6, the same comparisons are displayed for mesh refinement and error estimation on the field, using the errors (23) and (21), respectively.

DISCUSSION AND CONCLUSIONS

The extensive work on the "optimization" of the σ and κ refinement parameters of eqs. (10) and (8), even if by necessity semi-qualitative, has brought to the conclusion that a value around .5 of the σ parameter generates the most "balanced" meshes for all refinement methods, both for potential and field errors. As far as κ is concerned, it can be stated that for "field oriented" meshes again a value of .5 has proven to be the most adequate solution, for all methods that allow its variations. The situation is more dependent on the specific error estimate used for "potential oriented" meshes. As for the convergence of adaptive meshing to the true solution, all methods tested (with the exception of the "field difference" one, estimating the field error only) provide a good convergence, as can be seen in Fig. 3, giving error indicators on the potential, as defined by equation (22), lower than two per cent in the high field region. The relative merits of the various procedures for the solution of potential problems should then be evaluated on the basis of their computational effort. This comparison favours significantly the

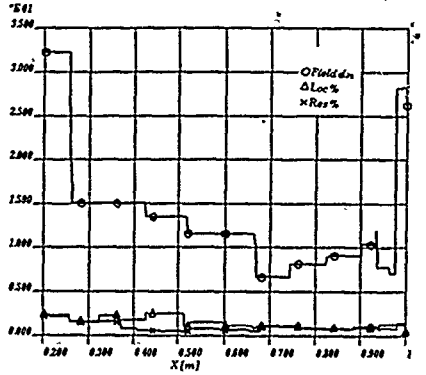


Fig. 3 - "True" relative errors on the potential, along A-B line, on the solution obtained using the various methods as refinement indicators.

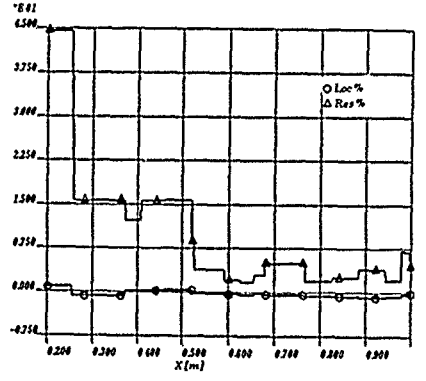


Fig. 4 - Difference between the "true" and the estimated errors (22) and (20) for the solution of potential problems.

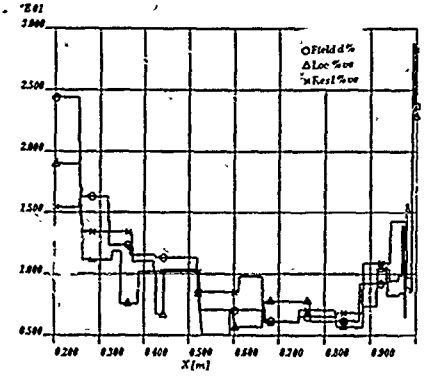


Fig. 5 - "True" relative errors on the field, along the A-B line of Fig. 1, on the solution obtained using the various methods as refinement indicators.

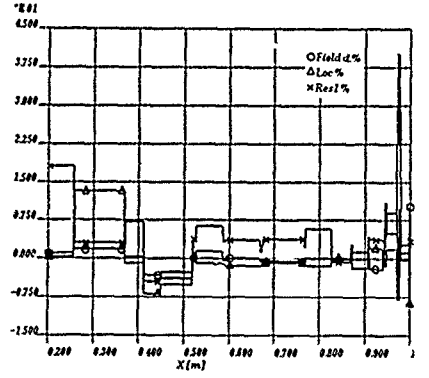


Fig. 6 - Difference between the "true" and the estimated errors (23) and (21) for the solution of field problems

"extended complete residual" formulations with the "optimal" value of κ being egual to zero, meaning an evaluation of the refinement estimator (8) only on the potential. The convergence to the true solution for the fields also shows a similar behaviour for all methods tested, with computational time slightly favouring the "local error problem" approach. Finally, regarding the final error estimation of the potential for the various methods, it can be observed in Fig. 4 that the best results are given so far by the "local error problem" algorithm. As far as the final error estimation of the field is concerned, it can be seen from Fig. 6 that all the methods seem to provide reasonably adequate estimates, even if none of the methods implemented has provided a uniform upper bound to the error. The activity is now continuing to verify the consistency of the above results in different cases, to improve mesh refinement convergence criteria and to try to define accelerated convergence strategies.

REFERENCE

[1] R. E. Bank and A. Weiser: " Some a-posteriori error estimators for elliptic partial differential equations", Mathematics of Computation, Vol. 44, pp. 283-301, Apr. 1985.

[2] I. Babuska and W. C. Rheinboldt: " Error estimates for adaptive finite element computations", SIAm J. Numer. Anal., Vol. 15, pp. 736-754, Aug. 1978.

[3] P. Fernandes, P. Girdinio, P. Molfino, M. Repetto: "Local error estimates for adaptive mesh refinement", IEEE Trans. on Mag., Vol. 24, pp. 299-302, Jan. 1988.

[4] P. Fernandes, P. Girdinio, P. Molfino, G. Molinari, M. Repetto: "A comparison of adaptive strategies for mesh refinement based on a posteriori local error estimation procedures", IEEE Trans. on Mag., Vol. 26, pp. 795-798, Jan. 1990.

[5] P. Fernandes, P. Girdinio, G. Molinari, M. Repetto: "Local error estimation procedures as refinement indicator in adaptive meshing", Fourth IEEE Conference on Electromagnetic Field Computation, Toronto, Oct. 22-24, 1990, paper EB-03.

[6] Z. J. Cendes and D. N. Shenton: "Adaptive mesh refinement in the finite element computation of magnetic fields", IEEE Trans. on Mag., Vol. MAG-21, pp. 1811-1816, Sept. 1985.

[7] D. W. Kelly, J. P. De S. R. Gago and O. C. Zienkiewicz: "A-posteriori error analysis and adaptive processes in the finite element method: Part I - error analysis", Inter. Journ. for Numerical Meth. in Engineering, Vol. 19, pp. 1593-1619, 1983.

[8] G. Molinari et al.:"A modular finite element package for research in electromagnetic analysis developed in a group of italian universities", Proc. of BISEF '88, Beijing, China, Sept. 1988.

RELATIONSHIPS BETWEEN DUAL ENERGY FORMULATIONS IN ELECTROMAGNETISM.

J Penman
Department of Engineering, University of Aberdeen, UK

Abstract:

It is the intention of this contribution to show how different ways of constructing dual energy forms for problems in electromagnetism can be related to each other, and then used to provide error bounded solutions for this class of problem. The first approach is based upon the principles of Hamilton and Toupin and the use of Legendre transformations, which can be related to the method of Lagrange multipliers in constrained optimization. The second technique stems from a geometrical interpretation in Hilbert space called the hypercircle method, first introduced by Prager and Synge [1]. Both can be shown to provide the same formulations as those achieved using direct integration and the methods of functional analysis. This allows a general structure for electromagnetic field problems to be developed which leads to the easy selection of appropriate energy functionals.

1. Hamilton's and Toupin's principle.

These principles has been used by several authors to obtain complementary variational forms for engineering field problems, in particular Hammond & Penman [2], used the analogy betwen analytical mechanics and electromagnetism to find complementary forms for electromagnetic field problems. To illustrate, the canonical set of equations for electrostatics, in a domain Ω is:

$-\nabla\phi = E$, with $\phi = g$ on $\partial\Omega_1$; $\quad \epsilon E = D$; $\quad \nabla.D = \rho$, with $n\,D = h$ on $\partial\Omega_2$.

Application of Hamilton's principle leads to the energy statement,

$$\theta(\phi) = {}^1\!/_2\epsilon(\nabla\phi|-\nabla\phi)_\Omega - (\rho|\phi)_\Omega + (h|\phi)_{\partial\Omega}$$

whilst Toupin's principle, by letting $D = D_S + \nabla\times C$ gives the complementary energy form,

$$\Xi(C) = {}^1\!/_2({}^{-1}\!/_\epsilon(D_S + \nabla\times C)|(D_S + \nabla\times C))_\Omega - (gn|(D_S + \nabla\times C))_{\partial\Omega}$$

These functionals are the primal complementary pair and may be extremized to provide bounded solutions to the the primal canonical equation. It is also possible to develop a dual system and a corresponding pair of complementary functionals, which also provide bounds. A closed chain of Legendre transformations can be used to move between the various forms,[3].

2. The hypercircle method.

In [1] it is shown how the solution to the appropriate boundary value problem can be constrained to lie on the hypercircle representing the intersection of a vector space hypersphere and a hyperplane. In the electrostatic example referred to earlier D is a function in the space V, in which distance is measured by the energy norm (E|D), where $E={}^1\!/_\epsilon D$. If two subspaces of V, V' and V", exist where V' contains functions D' such that $D = \epsilon E$ and $\nabla.D' = \rho$, and V" contains functions D" such that $D" = \epsilon E"$ and $\nabla\times E" = 0$, then the intersection of V' and V" represents the exact function $D = \epsilon E$, which also satisfies $\nabla.D = \rho$ and $\nabla\times E = 0$. The energy norms in these spaces are also the primal complementary pair given above.

3. The direct method.

More generally, the partial differential equations of electromagnetism often have the form, $T^a M(Tu+v_s) = p$, in Ω, with $Bu = g$ on $\partial\Omega_1$ and $B^*w = h$ on $\partial\Omega_2$, where $w = M(Tu+v_s)$. Here, T^a is the adjoint of T, M is self-adjoint and invertible, u and $p \in U$, and w and $v \in V$. Using the methods described by Vainberg, [4] one can generate the functional,

$$\Phi(u,v,w) = ({}^1\!/_2T^a\text{-}p|u)_U + ({}^1\!/_2B^*w\text{-}h|u)_{\partial\Omega} + {}^1\!/_2(Mv\text{-}w|v)_V + ({}^1\!/_2Tu\text{-}{}^1\!/_2v+v_s|w)_V + ({}^1\!/_2Bu\text{-}g|w)_{\partial\Omega}.$$

Using a generalised form of Greens theorem this functional can be reduced to two different forms, namely,

$$\theta(u) = {}^1\!/_2(Tu+v_s|M(Tu+v_s))_U - (p|u)_U - (h|u)_{\partial\Omega}, \text{ and}$$

$$\Xi(w) = ({}^{-1}\!/_2M^{-1}w+v_s|w)_V - (g|w)_{\partial\Omega}.$$

Reduced to the same nomenclature these are exactly the fuctionals given above for the electrostatic system. They, have much wider applicability however, and can be used to provide error bounded pairs of functionals for magnetic and electromagnetic field problems too. The finite element method can be used to extremize them to a prescribed level of accuracy.

4. An example from the magnetic field

In section 3, above, the solution to $T^2M(Tu+v_s) = p$ was sought by extremizing the functionals, $\theta(u)$ and $\Xi(w)$. This equation can be expressed in canonical form as,

$Tu+v_s=v$, with $Bu = g$ on $\partial\Omega_1$
$Mv=w$ in Ω
$T^2w=p$ with $B^*w = h$ on $\partial\Omega_2$.

For the magnetic field the equivalent set is,

$\nabla\times A=B$ with $A=g$ on $\partial\Omega_1$
$H=1/\mu(B)$ in Ω
$\nabla\times H=J$ with $n\times H=h$ on $\partial\Omega_2$,

and the corresponding functionals are,

$$\theta(A) = \tfrac{1}{2}(\nabla\times A|1/\mu(\nabla\times A))_\Omega - (J|A)_\Omega + (h|A)_{\partial\Omega}$$

and,

$$\Xi(H) = \tfrac{1}{2}(H|\mu H)_\Omega - (g|n\times H)_{\partial\Omega}.$$

To illustrate the bounded nature of this complementary pair the simple example of figure 1 is solved, for a several different numbers of degrees of freedom, using the finite element technique. The convergence in energy, from above and below the true value is shown in figure 2, and the magnetic flux density distributions calculated from A and H, respectively is given in figures 3 and 4. The expected similarity of the two fields is apparent.

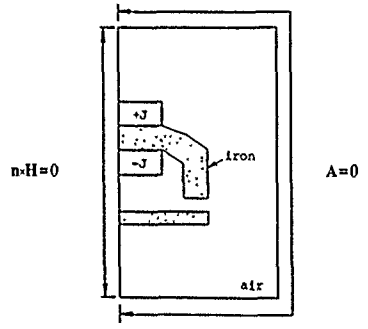


Figure 1: A magnetic field problem.

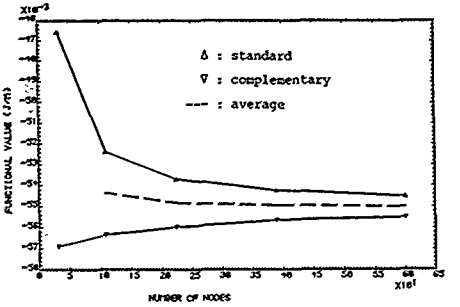


Figure 2: Error bounded convergence.

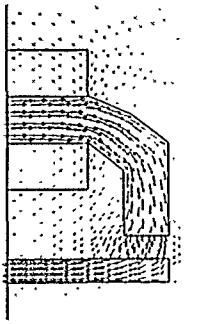


Figure 3: B from $\theta(A)$.

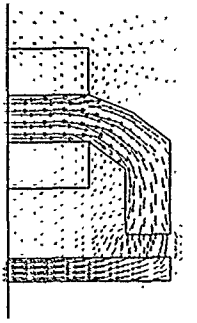


Figure 4: B from $\Xi(H)$

5. References.

1. Synge JL, *The hypercircle in mathematical physics*, Pub CUP, 1957

2. Hammond P, Penman J, *Calculation of eddy currents by dual energy methods*, ProcIEE, 125, 1978, pp701-708

3. Sewell MJ, *On dual approximation principles in continuum mechanics*, Trans Royal Soc, 265, 1969, pp319-350

4. Vainberg MM, *Variational methods in the theory of non-linear operators*, Pub J Wiley, 1973.

SOLVING MAXWELL'S EQUATIONS USING VECTOR POTENTIALS FOR THE FULL FREQUENCY SPECTRUM

C R I Emson and C W Trowbridge
Vector Fields Ltd., 24 Bankside, Kidlington, Oxon OX5 1JE, UK

Abstract – The aim of the present paper is to present a general method using vector potentials, and show its applicability to electromagnetic field problems over the entire frequency spectrum. Results will be given for a range of problems using a general purpose computer code based on the finite element method.

Introduction

A great deal of work has been carried out recently in defining robust algorithms for 3D eddy current analysis, producing unique and accurate solutions. The problem of gauging vector solutions has therefore been carefully studied, resulting in a number of viable formulations for this 'low' frequency limit of Maxwell's equations [1,2].

The formulation used in this paper is based on the Lorentz gauge, which is extended to solve the full set of Maxwell's equations. This can then lead to either an eigenvalue problem (in which conductors are assumed ideal, and are represented as boundary conditions), or to a full deterministic problem (when lossy materials are included).

Low Frequency Limit Of Maxwell's Equations

The governing equation in terms of the magnetic vector and electric scalar potential is well known for the low frequency case,

$$\nabla \times \frac{1}{\mu} \nabla \times \mathbf{A} = -\sigma \frac{\partial \mathbf{A}}{\partial t} - \sigma \nabla V + \sigma(\mathbf{u} \times \nabla \times \mathbf{A}) \qquad (1)$$

along with the required condition $\nabla \cdot \mathbf{J} = 0$

$$\nabla \cdot \left(\sigma \frac{\partial \mathbf{A}}{\partial t} + \sigma \nabla V - \sigma \mathbf{u} \times \nabla \times \mathbf{A} \right) = 0 \qquad (2)$$

The use of the Lorentz gauge $\nabla \cdot \mathbf{A} = -\mu\sigma V$ to separate A and V equations is classical, and it has recently been shown that the uniqueness proofs of equations (1) and (2) require a further condition, for example $\mathbf{A} \cdot \mathbf{n} = \beta V$ on conductor boundaries (where β is an arbitary constant). This results in a form of eqn (1) independent of V, with the condition $\mathbf{A} \cdot \mathbf{n} = 0$ included in a weak form. The current can later be found by solving eqn (2), where A is the value obtained from above.

This formulation is also applicable for DC problems (since the gauge is independent of frequency), and allows vector potentials to be used in non-conducting regions (with the gauge effectively becoming the Coulomb gauge in those volumes). Transient solutions to the above equation can also be obtained using a time integration scheme. Results for a transient problem are shown in Figures 1 and 2, showing half the geometry (see Problem 14 of the TEAM Workshop [3]) and the time variation of power loss in the casing.

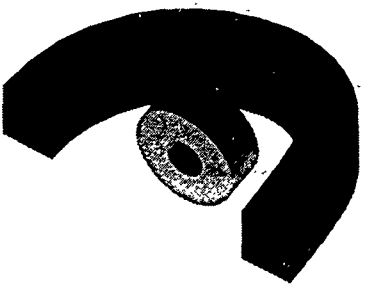


Figure 1: *Geometry of a transient eddy current field problem*

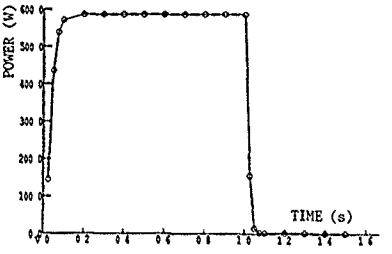


Figure 2: *Solution to transient eddy current field problem*

The effects of velocity can be included also, leading to the extra terms in eqn (2) involving the velocity u. The resulting equation is complex for time harmonic problems, but is no longer symmetric. A conjugate gradient squared algorithm is used therefore to solve the linear equations.

Formulation For The Full Frequency Spectrum

When the frequency is high, it is necessary to include the effects of displacement currents. The full system shown in equation (3) must now be solved (neglecting velocity terms).

$$\nabla \times \frac{1}{\mu} \nabla \times \mathbf{A} = -\sigma \frac{\partial \mathbf{A}}{\partial t} - \varepsilon \frac{\partial^2 \mathbf{A}}{\partial t^2} - \sigma \nabla V - \varepsilon \frac{\partial \nabla V}{\partial t} \qquad (3)$$

The full form of the Lorentz gauge is used $\nabla \cdot \mathbf{A} = -\mu\sigma V - \mu\varepsilon \frac{\partial V}{\partial t}$ and when substituted into equation (3) leads to the new governing equation (after having applied the Galerkin procedure)

$$\int \nabla \times \mathbf{W} \frac{1}{\mu} \nabla \times \mathbf{A}\, d\Omega + \int \nabla \cdot \mathbf{W} \frac{1}{\mu} \nabla \cdot \mathbf{A}\, d\Omega - i\omega \int \sigma \mathbf{W} \cdot \mathbf{A}\, d\Omega - \omega^2 \int \varepsilon \mathbf{W} \cdot \mathbf{A}\, d\Omega + \int_\Gamma \frac{1}{\beta} \mathbf{A} \cdot \mathbf{n}\, \mathbf{W} \cdot \mathbf{n}\, d\Gamma = 0 \qquad (4)$$

This is then solved using a standard complex ICCG solver. To be able to recover the electric field (including the losses), the solution for V must be obtained. This is achieved by solving the secondary equation $\nabla \cdot \mathbf{J} + \nabla \cdot \frac{\partial \mathbf{D}}{\partial t} = 0$.

If the frequency is sufficiently high, the conductors can be regarded as ideal In this case, the terms involving σ can be neglected, and conductors are modelled using boundary conditions alone. This leads to a real set of equations. Alternatively, the equations can be structured as an eigenvalue problem, of the form $\mathbf{Ku} = \omega^2 \mathbf{Mu}$, which can be solved to find the resonant frequencies of the system.

A typical example is shown below shows the standing waves generated by a short circuit termination on a rectangular waveguide at two different frequencies, one above cutoff and the other below (showing the evanescent nature of the solution).

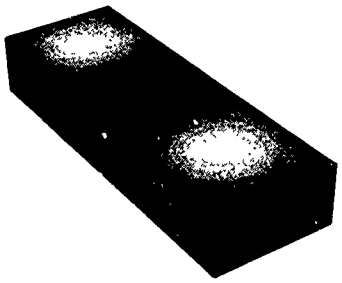


Figure 3: *Field E_y in short circuit waveguide at 10 Ghz (above cutoff)*

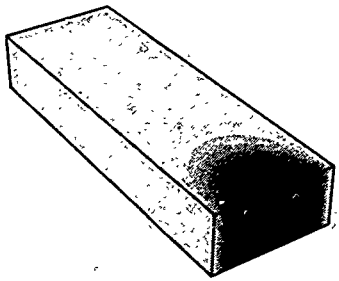


Figure 4: *Field E_y in short circuit waveguide at 1 GHz (below cutoff)*

Conclusions

A formulation, similar to the low frequency eddy current formulation presented elsewhere, has been shown to be extendable to solve high frequency applications. An eigenvalue solver has been developed, and results have shown that the approach is viable, and does not produce any spurious modes.

The full set of Maxwell's equations have also been implemented. The resulting equations are no longer of an eigenvalue nature, and a deterministic solution method is required. This will then enable solutions to problems in which lossy materials are also present, in addition to lossless dielectric and permeable material types.

References

[1] O. Biro and K. Preis, "Finite element analysis of 3-D eddy currents," *IEEE Transactions on Magnetics*, vol. MAG-26, p. 418, March 1990.

[2] C F Bryant, C R I Emson and C W Trowbridge, "A general purpose 3-D formulation for eddy currents using the Lorentz gauge," *IEEE Transactions on Magnetics*, vol. MAG-26, p 2373, September 1990.

[3] "Proceedings of the Oxford TEAM workshop, 23-25 april 1990," Tech. Rep., ISPRA, 1990.

Magnetostatic conservation principles and mixed FEM.

S Polak

April 9, 1991

Abstract

In this paper we consider the mixed FEM in connection with conservation principles for magnetostatic problems. The usual FEM can be viewed as giving a minimal energy in the basis function space under consideration It does not satisfy a discrete flux conservation principle. Mixed FEM's have the property that they satisfy such principles. The mixed FEM is explained with the help of a simple Poisson problem. Then we show the lack of conservation for the FEM. Several formulations of the magnetostatic problem are discretised with the mixed FEM and conservation principles found. Also normal and tangential component continuity properties of vector quantities involved, a strong point of the mixed FEM, are discussed.

1 Introduction.

In this paper we investigate the conservation principles and element interface conditions that are of interest for magnetostatic problems in connection with mixed FEM. In section 2 we describe the mixed finite element method for the simplest Poisson problem. In section 3 we show with a simple example that the finite element method does not have the conservation property under consideration. In section 4 conservation principles for magnetostatic problems are discussed in connection with mixed FEM discretisations The term "exact" will be used to indicate that apart from rounding errors two numbers are equal, so the fact that equality is not influenced by discretisation.

2 Mixed Finite Elements for the Poisson problem.

The mixed finite element method is discussed here only with respect to applications For more basic mathematical considerations [1] may be consulted In this section we explain the mixed FEM for the Poisson problem $\Delta u = \rho$, on a region $\Omega \subset \Re^2$ with boundary Γ, and u given on Γ. It enables us to explain the essence without unnecessary complications. The basic idea is to split the equation in two first order equations

$$\sigma = \nabla u \tag{1}$$
$$\nabla \cdot \sigma = \rho \tag{2}$$

The mixed FEM formulation for this problem then is

$$(\sigma, \tau) + (u, \nabla \cdot \tau) = r_1 \tag{3}$$
$$(\nabla \cdot \sigma, \phi) = r_2 \tag{4}$$

$\forall \tau \in W_h$ and $\forall \phi \in V_h$, where $(\ ,\ )$ is defined by $\int\int_\Omega$ and r_1 stems from the boundary conditions, $W_h \subset H(div;\Omega)$ and $V_h \subset L_2(\Omega)$ In practice the simplest basis functions for a triangular mesh are given by the u, ϕ constant on each triangle and σ, τ piece wise linear vector functions with continuous normal components on the edges. These are e g. the lowest order Raviart Thomas spaces, [2]. A basis for the vector space is given by $\alpha_i(x - x_i, y - y_i)^T$ where α_i is the normalizing factor. The set of equations stemming from this discretisation can be represented by

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \dot{\sigma} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} \dot{r}_1 \\ \dot{r}_2 \end{pmatrix} \tag{5}$$

with $A = ((\tau_i, \tau_j)), B = ((\phi_i, \nabla \cdot \tau_j))$. The τ_i are a basis for V_h and ϕ_i a basis for W_h. The unknown are the normal components at the element edges and the constants per element. Here we have not performed a partial integration on the $\nabla\cdot$ applied to the vector unknown as is done in ordinary FEM This is the basis of the conservation principle that we will come back to later. Contrary to what we would expect from the differentiability requirements displayed by the operational formulation, the potentials are represented as piece wise constants, the vector unknowns as linear functions. To solve equation 5 for the potential values after elimination of the vectors one encounters the matrix $B^T A^{-1} B$. This is not an M matrix and it is also dense. One may wonder whether there is any advantage in not eliminating the matrix A and solving with e g. a preconditioned iterative method. I have no knowledge of any special properties that would make this advantageous. The usual procedure is [3] the introduction of discontinuities in the edge normal components of the vector unknown in the space of vector functions The continuity is enforced by adding extra equations. The extra equation number is compensated by the fact that now there are two unknown per edge. There still are the same number of unknown and equations. The equation 5 now is replaced by

$$\int\int_\Omega \sigma \cdot \tau + \sum_T \int\int_T u \nabla \cdot \tau - \sum_T \int_{\partial T} \lambda \tau \cdot dn = 0 \tag{6}$$
$$\int\int_\Omega (\nabla \cdot \sigma) \phi do = \int\int_\Omega \rho \phi do \tag{7}$$
$$\sum_T \int_{\partial T} \xi \sigma \cdot dn = 0 \tag{8}$$

$\forall \tau \in \vec{W}_h, \forall \phi \in V_h$ and $\forall \xi \in \Lambda_{h,0}$ with $\vec{W}$ the space of linear functions on each triangle, $\Lambda_{h,\xi}$ the space of edge valued functions that are equal to ξ on Γ. This set of equations can be represented by

$$\begin{pmatrix} A & B & C \\ B^T & 0 & 0 \\ C^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma^* \\ u^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} r_1^* \\ r_2^* \\ r_3^* \end{pmatrix} \tag{9}$$

where σ^*, u^*, λ^* and r_i^* are the logical replacements of the variables σ, u and r as used in equation 5. It is shown in [1] that the matrix in equation 2 is a row diagonally dominant M-matrix for non obtuse triangles. The solution obtained with both formulations is exactly the same. It is clear that, for the same number of triangles we have more unknown in the mixed FEM than in an ordinary FEM. For each triangle there are three unknown edge variables, one unknown potential value and three unknown vector normal components. Per two triangles there is one edge variable λ. So for N triangles approximately (boundary conditions) 5 5N unknown. However we may eliminate the constant potential values and the normal components of the vector unknown The system for the edge variables still is sparse, with in two dimensions only a five point coupling. This implies that the matrix has less nonzero values on each row then usual with FEM. The number of unknown now is approximately 1 5N. For FEM we can not exactly make a similar calculation because the number of unknown depends on the number of triangles coming together in one vertex. However if we assume an average of six triangles per vertex we have approximately 3N/6 unknown. The conclusion is that the number of unknown is more but the number of nonzeroes is less in mixed FEM than in ordinary FEM. There is no evidence known to me clearly showing the properties of the two types of matrices with respect to preconditioned iterative methods. It also is unknown to me whether the same number of triangles is needed for equal accuracy.

3 Conservation principles and FEM.

The FEM often is derived by requiring an energy functional to have a minimum value over a certain finite dimensional space The continuous solution often satisfies a conservation principle at the same time So for 2 which we may interpret as an electrostatic equation relating the displacement vector $D = \nabla u$ to the charge density ρ the conservation principle states that the charge in a volume equals the total displacement over the boundary of the volume A simple example shows that within the representation and spaces used in the FEM this is not exactly the case for the FEM solution Consider the example

$u_{xx} = 1, u(0) = u(3) = 0$ on the segment $[0,3]$. For the conservation principle we should find $u_x(3) - u_x(0) = \int_0^3 1 dx = 3$. However the FEM solution with one unknown at $x=1$ and linear basis functions will give 2.25 instead of 3. For the mixed FEM discussed in section 2 the conservation is guaranteed by the fact that one of the two (or three) basic equations is

$$\int\int(\nabla \cdot \sigma - \rho)\phi d\sigma = 0 \tag{10}$$

$\forall \phi$ with ϕ constant per triangle. We therefore may take $\phi = 1$ on a triangle. Using the Gauss theorem immediately gives the conservation principle. It should be stressed that this is mainly due to the fact that we did not use a partial integration to derive in the first place.

4 Conservation principles for magnetostatic problems.

First we consider the usual magnetostatic vector potential formulation. So

$$\nabla \times \nu \nabla \times A = J \tag{11}$$

In two dimensions this is exactly equation 2. However the interpretation of the conservation principle is different. We have $\nabla \times A \cdot \nabla A_z = 0$ and $\|\nabla \times A\| = \|\nabla A_z\|$ where $A = (0,0,A_z)$. Therefore the conservation principle is simply $\oint H \cdot ds = \int\int J d\sigma$. Two facts are interesting, the conservation is per arbitrary group of triangles exact and for any element edge there is point wise tangential component continuity of H. It should also be pointed out that the calassical "spurious sources" are impossible with the mixed FEM. In three dimensions the situation is more complicated because as is well known [4] a gauge condition is needed to ensure uniqueness for the vector potential. A possibility is the inclusion of the gauge by using a $\nabla \cdot \nabla$ formulation as e.g. discussed in [5]. I suspect that also here a conservation principle can be found, but have to investigate this further. The other formulation that is of interest is [6]:

$$\nabla \cdot \mu(\nabla\phi - H_c) = 0 \tag{12}$$

where H_c is as defined in [6]. The first order equations are

$$\nu B = (\nabla\phi - H_c) \tag{13}$$

$$\nabla \cdot B = 0 \tag{14}$$

where the way μ is treated is important. Then the mixed FEM with normal component continuity is

$$(\nu B, \tau) + (\phi, \nabla \cdot \tau) - (H_c, \tau) = 0 \tag{15}$$

$$(\nabla \cdot B, u) = 0 \tag{16}$$

$\forall \tau \in W_h$ and $\forall u \in V_h$ and the equations with normal component discontinuity for the basis functions (not for the solution!).

$$\int\int_\Omega \nu B \cdot \tau + \sum_T \int\int_T (\phi \nabla - H_c)\tau$$

$$-\sum_T \int\int_{\partial T} \lambda \tau \cdot dn = 0 \tag{17}$$

$$\int\int_\Omega (\nabla \cdot B)\phi d\sigma = 0 \tag{18}$$

$$\sum_T \int\int_{\partial T} \xi B \cdot dn = 0 \tag{19}$$

$\forall \tau \in W_h, \forall \phi \in V_h$ and $\forall \xi \in \Lambda_{h,\rho}$. Therefore the conservation principle here is

$$\int\int_\Gamma B \cdot dn = 0 \tag{20}$$

for Γ the boundary of any set of elements. As usual in the mixed FEM we also find normal component continuity, point wise on any element edge. Also when the combination of a reduced and a total potential [6] are used the formulation automatically gives a pointwise B normal component continuity at the interface. For Eddy current problems the same discretisation principle can be applied but the simplest basis functions no longer give an M-matrix. Solutions for this problem are discussed in [7] and [8].

5. Concluding remarks.

From the considerations given in this paper it is clear that conservation principles are found when using the mixed FEM for magnetostatic problems, similar to those found for other problems. The efficiency of the method w.r.t. the FEM can not be judged from theoretical considerations as there are on the one hand more unknown while on the other hand the matrix is more sparse. On top of this it is not clear how the number of necessary triangles differs from the number needed for the FEM. Practical experience with mixed FEM for magnetostatic problems would be very interesting.

References

[1] F.Brezzi.'On the existence, uniqueness and approximation of saddle point problems arising from lagrangian multipliers' Revue Francaise d'Automatique et Recherche Operationelle, août1974,R-2.

[2] P.A.Raviart and J.M.Thomas. A mixed Finite Element Method for second order elliptic problems', Mathematical aspects of the Finite Element Method. Lecture Notes in Math 606,292-315, Springer 1977.

[3] N X Fraeijs de Veubeke.'Displacement and equilibrium models in the finite element method'. Stress Analysis, O.C.Zienckiewicz and G Hollister, eds Wiley New York,1965.

[4] S J.Polak,et al ,'A new 3-D eddy current model', IEEE Trans Mag. Vol.19, No.6 1983,pp 2447-2449.

[5] T.Morisue,'A new formulation of the magnetic vector potential method in 3D multiply connected regions', IEEE Trans. Mag., Vol.24, No.1,1988, pp 110-113.

[6] J.Simkin and C.W. Trowbridge'On the use of the total scalar potential in the numerical solution of field problems in electromagnetics.' IJNME,vol. 14,423-440(1979).

[7] S.J.Polak'Mixed FEM for $\Delta u = \alpha u$' International series of numerical mathematics', vol. 93,1990, Birkhauser Verlag, Basel.

[8] Luisa Donatella Marini and Paola Pietra'New mixed finite element schemes for continuity equations', Compel, vol 9, number 4

OPEN BOUNDARY PROBLEMS IN ELECTROMAGNETIC CALCULATIONS

by Gérard Meunier (1) and Xavier Brunotte (1, 2)
(1) Laboratoire d'Electrotechnique de Grenoble (URA 355 CNRS), INPG
(2) Laboratoire de Magnétisme du Navire, INPG
ENSIEG, Domaine Universitaire BP 46, 38 402 Saint Martin d'Hères, France

ABSTRACT

This paper presents a finite element technique for the computation of open boundary problems using transformations. The principle of the method is brifly developed. The efficiency and the easiness of the implementation renders this method very interesting. An example of an magnetodynamic application is presented.

INTRODUCTION

Unbounded problems in electromagnetism science are encountered in a great number of applications. The resolution of such problems is not obvious using the finite element method (FEM). Indeed the magnetic or electric fields far away from the structure can not be neglected.

In order to overcome this limitation of the FEM, several techniques have been proposed by differents authors[1,2]. According to P.Bettes[2], they might be separated into five different groups:

- the method of truncation wich consist to approximate the infinite domain by a closed domain, which should be sufficiently large
- the method called "Ballooning" developped by P.P. Silvester and al[3]
- the "infinite" or "mapped" elements[4]
- the use of transformations reducing an open infinite region to a closed one[5]
- the methods coupling FEM with an analytical or numerical method (BEM)[6,7]

Among these methods, we have used particulary those involving transformations. The theory applied is to reduce an open infinite region to a closed finite one. Up to now, the transformation used (inverse tranformation like 1/r) were conformal mappings logically limited to applications in 2D cartesian coordinates from a geometrical point of view and to the Laplace equation from a physical point of view (the Laplace equation is invariant to conformal mapping). In previous papers, J.F. Imhoff [8,9]demonstrated that these transformations need not to be conformal and therefore may be applied to problems other than the Laplace equation in 2D cartesian coordinates. He has particulary extented the possibilities of inverse transformation applications to axisymmetrical and 3D problems and has developed a transformation which even in 2D is not conformal but whose introduction into a 3D finite element software is easier than that one of an inverse transformation. These works being very interesting, we have been searching for new transformations and introduce them into the software FLUX3D with regard to a great efficiency.

PRINCIPLE OF THE METHOD

The principle consists in considering that the entire 2D or 3D space E is the sum of a closed internal subdomain E_{int} (treated by the FEM) and of an open boundary external subdomain E_{ext}.

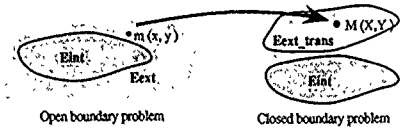


Figure 1

In order to take into account the external domain, we map it into a closed domain E_{ext_trans} by means of a bijective spatial transformation as it appears in figure 1. This domain E_{ext_trans} is meshed, and used for solving the problem. The equations of the physical problem are, of course modified by the transformation. For electromagnetic problems, the alterations of the equations are obvious. For example, using Galerkin method for the magnetostatic scalar model, we usually have to compute :

$$\iiint_{Ext} \mu \left[\frac{\partial w_i}{\partial x}\right].\left[\frac{\partial V}{\partial x}\right]^T dx =$$

$$\iiint_{Eext\_trans} \mu \left[\frac{\partial W_i}{\partial X}\right].\left[\frac{\partial X}{\partial x}\right].\left[\frac{\partial X}{\partial x}\right]^T.\left[\frac{\partial V_j}{\partial X}\right]^T . det\left[\frac{\partial X}{\partial x}\right]^{-1} dX$$

where x are the coordinates in the real space E_{ext}, X the coordinates in the image space E_{ext_trans}, $[\partial x/\partial X]$ is the Jacobian matrix of the transformation, W_i are the usual finite element functions in the mapped domain and w_i are weighting Galerkin functions, which are known thanks to their images W_i.

From all the transformations we tried, we have kept the spherical shell transformation[9] and the parallelepipedic transformation[10]. The way we implement these transformations lead to the creation of an efficient tool. The modification of the integration procedure is made by subroutines wich carry out the modification of the Galerkin functions due to the transformation.The post-procesing (fields computation in the infinite region for example) is made by the same way using the same subroutines.

EXAMPLE

The Figure 2 presents a magnetodynamic problem: computation of eddy currents in a thin sheet conductor induced by a exterior sinusoïdal field (0.1 Hz). The sheet conductor is modelled by a A-V formulation applied to surface element. The exterior uniform field is simulated using the reduced scalar potential and the infinite domain is modeled by a parallelipipedic transformation. Figure 2 shows eddy currents in the sheet and reduced scalar equipotentials in the air.

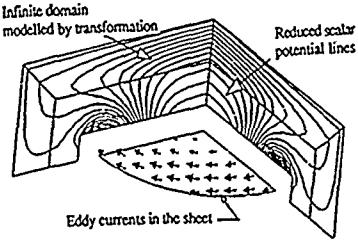


Figure 2

CONCLUSION

The method used gives very good results for the exterior region and is very cheap compared to the "Balloning" method or hybrid method (FEM-BEM). The employed transformations seem to be similar to the transformation develloped by Zienkiewicz, Emson and Bettes[4] which introduced the mapped element.They give the same relative and absolute error. Howewer their transformation can only simulate the exterior region by using one element. With the method presented in this paper we make a global definition of the transformation which allows to mesh the exterior region as thickly as necessary in ordrer to improve the solution.

REFERENCES

[1]P. Bettes, *"Finite element modelling of exterior electromagnetic problems"* , IEEE TransMag, Vol. 24, NO. 1, Jan. 88

[2]C.R.I. Emson, *"Methods for the solution of open-boundary electromagnetic-field problems"*, IEE Proc., Vol. 135, Pt.A, N°3, March 1988

[3]P.P. Silvester, D.A. Lowther, C.J. Carpenter, E.A. Wyatt, *"Exteroir finite elements for 2-dimensional exterior field problems"*, Proc. IEE 124, p. 1267, December1977

[4]O.C. Zienkiewicz, C. Emson, P. Bettes *"A novel boundary infinite element"*, International Journal for Numerical Methods in Engineering, Vol. 19, pp 393-404, 1983

[5]D.A. Lowther, E. Freeman, B. Forghani, *'A sparse matrix open boundary method for finite element analysis"*, IEE Transactions on Magnetics, Vol. 25, N° 4, July 1989

[6]O. C. Zienckiwicz, D.W. Kelly, *"The coupling of finite element and boundary solution procedures"*, International Journal for Numerical Methods in Engineering, No 11, PP 355-375, 1977

[7]G. Meunier, J.L. Coulomb, S. Salon, L. Krähenbühl, *"Hybrid finite element solutions for three dimensionnal scalar potential problems"*, IEEE Trans-Mag Vol-22, n° 5 p1040, Septembre 1986

[8] J.F. Imhoff, G. Meunier, J. C. Sabonnadière : "Finite Element Modeling of Open Boundary Problems" COMPUMAG, Tokyo september 1989, IEEE Transactions on Magnetics.

[9]J.F. Imhoff, G. Meunier, X. Brunotte, J.C. Sabonnadière: *"An original solution for unbounded electromagnetic 2D and 3D problems throughout the finite element method"*, INTERMAG 90, Brighton, IEEE Transactions on Magnetics N°26, Septembre 1990, pp. 2196-2199

[10]X.Brunotte, G. Meunier, X. J.F. Imhoff: *"AFinite element modeling of unbounded problems using transformations: a rigourous, powerfull and easy solution"*, COPUMAG 91.

DIFFERENT STRATEGIES IN THE OPTIMIZATION OF ELECTROMAGNETIC DEVICES

C.A.Magele, K.Preis, O.Biro, K.R.Richter

Graz University of Technology, Kopernikusgasse 24, A-8010 Graz, Austria

Abstract

The design of optimal geometric boundaries to achieve a prescribed behaviour of the electromagnetic field has been attracting more and more attention recently. This paper summarizes the application of "evolution strategies" [1,2,3] to find the global minimum of a given objective function These strategies utilize a simplified model of biological evolution. Another method used for global minimization is known as "simulated annealing" [4], a method based on simulation of the physical process of cooling solids to their lowest energy configuration. Once a solution in the neighbourhood of the optimum has been obtained by the methods mentioned above, a deterministic procedure, e.g. the Gauss-Seidel strategy or a higher order deterministic strategy can be used to "identify" the optimum in an efficient way. The determination of the magnetic field is carried out by a finite element calculation. The efficiency of the optimization strategies has been investigated using a simple two dimensional model with nonlinear material characteristics.

Introduction

Optimal design requires the solution of a synthesis problem The goal is to produce a given field by a sought geometry. This task, for example, is encountered if one has to design the magnet system of an NMR [5], a medical diagnostic equipment. Both deterministic and nondeterministic optimization methods need the successive recalculation of an electromagnetic field problem using each time a more or less different geometry. In order to keep the number of trials necessary to reach the optimal shape of the magnets or distribution of coils as low as possible, some strategy in performing the variations is required. Recently, evolution strategies have been introduced for the solution of electromagnetic optimization problems.

Evolution Strategies

Evolution strategies applied to electromagnetic problems [5,6] utilize a simplified model of biological evolution to arrive at an optimum of a given technical problem They are more likely to arrive at the global optimum than deterministic methods. To begin with, a set of variables which will be modified in the course of the optimization process must be set to their initial values. Common to all evolution strategies is the generation of one or more "descendants" from one or more "parents" by the addition of a random vector with normally distributed components and some mechanism to choose the descendant to "survive".

(1+1) Evolution Strategy

The simplest scheme simulating natural proceedings is the (1+1) strategy. In each generation a single descendant springs off a single parent. It replaces the parent only if it leads to a better quality of the objective function. This means, that no deterioration occurs from one generation to the next. After a certain number of generations the probability of arriving at a "better" descendant is evaluated and the standard deviation of the components of the random vector (mutual step width) is adaptively increased (too many successes) or decreased (too few successful attempts).

(1,6) Evolution Strategy

This higher order strategy makes use of some additional features of biological evolution, namely population, finite live span of an individual and genetically adjusted mutability 6 descendants are generated from one parent using two different values for the mutation step width. The new parent is selected from among the descendants only. The adaptive control of the mutation step width is achieved by the "survival" of the strategy variables of the chosen descendant. The flow chart of this scheme is shown in Fig.1 One major advantage of this and other higher order evolution strategies compared to the (1+1) strategy is the fact that sometimes a new accepted solution yields a worse quality than the current solution. The benefit of allowing deteriorating solutions is to help the search avoid becoming trapped in a local minimum of the objective function.

(10,100) Evolution Strategy

In this multi-membered scheme another aspect of evolution, recombination, can be introduced into the model for the generation of descendants. Every set of object variables has its corresponding set of strategy variables. 10 parents produce 100 descendants, allowing the 10 best ones to survive to become the new parent generation.

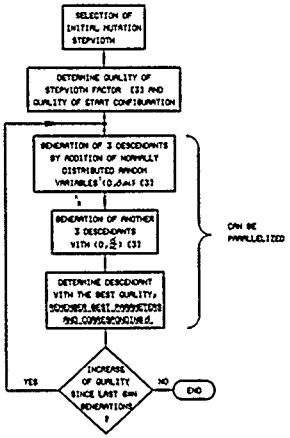


Fig.1 Flow-chart of the (1,6) evolution strategy

Simulated annealing

The name of this method derives from an analogy between simulating the annealing of solids and solving optimization problems. Starting from an initial state ("initial temperature") the trial variables are perturbed at random to a new state in the neighbourhood. This state corresponds to a new temperature, specified by a cooling schedule If the new state represents a reduction in the value of the objective function, then the transformation to the new state is accepted. Similar to higher order evolution strategies there is a chance that a state yielding a worse quality than the current one will be accepted. The acceptance probability function, based on the Boltzmann distribution, takes the form of $\exp(-\Delta Q/kT)$, where ΔQ is the increase in quality, T is the control paramter ("temperature") and k is an appropriate constant. The flow chart of this scheme is shown in Fig.2

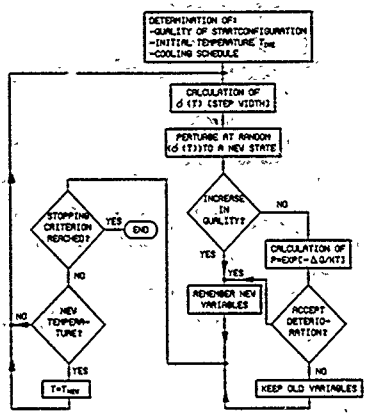


Fig.2 Flow-chart of "Simulated Annealing"

Using this method, a cooling schedule, the number of repetitions having a constant temperature and a function controlling the stepwidth are required and many different schemes to control the algorithm have been suggested [7].
Using for instance the (1,6) evolution strategy an initial stepwidth σ and a stepwidth factor α have to be specified only. This fact seems to make evolution strategies easier to handle.

Numerical Investigations

Three currently very important strategies, the (1+1) evolution strategy, the (1,6) evolution strategy and the simulated annealing method have been applied to a very simple 2D nonlinear magnetostatic optimization problem (Fig.3, initial geometry: planar pole $x_1=x_2=50$, $x_3=100$, dimensions in mm). The pole shape of a dipole magnet has been optimized (shimming problem), varying the trial variables x_1, x_2 and x_3. The goal was to achive a homogeneous field of 1 Tesla in the prescribed region under the pole.

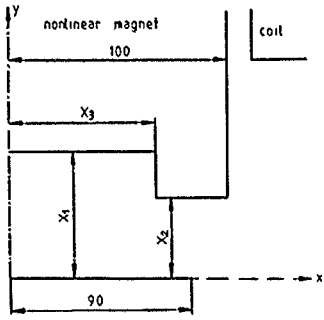


Fig.3 Shimming problem

Using the (1,6) evolution strategy the initial stepwidth σ was set to 10mm and the stepwidth factor α was set to 1,3 Using simulated annealing the temperature was assumed to decrease in a linear manner from T_{Start} = 1000 degrees to T_{End} = 0 degrees.

The stepwidth (σ_{Start} = 10 mm, σ_{End}= 1mm) was assumed to decrease in a exponential way, allowing more significant changes in geometry at higher temperatures:

$$\sigma(T) = \sigma_{End} + (\sigma_{Start} - \sigma_{End}) * (1 - \exp(-T/\tau)), \quad (1)$$

where τ =200 degrees
Every single test case was stopped after 100 finite element calculations. The best solutions arrived at the following values for x_1, x_2 and x_3 (dimensions in mm).

	x_1	x_2	x_3
(1+1)	51.67	31.77	90 7
(1,6)	51.85	28 8	91 87
SA	51.14	29 62	91 74

Fig.4 "optimized" paramters of the shimming problem

The distribution of the magnetic flux density in the region of interest obtained using the initial geometry and the above mentioned best solutions is shown in Fig.5. All methods under investigation lead to very good results. Using a finer discretization even better results will be obtained.

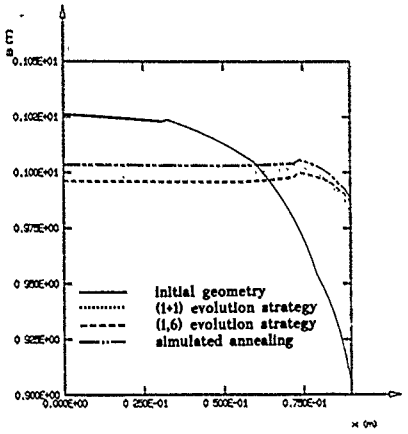


Fig.5 Distribution of the flux density in the region of interest

References

[1] I. Rechenberg, Evolutionsstrategie, Optimierug technischer Systeme nach Prinzipien der biologischen Evolution Stuttgart-Bad Cannstatd. Frommann-Holzboog, 1973

[2] H.P. S hwefel, Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie. Basel und Stuttgart: Birkhäuser Verlag, 1977

[3] A. Scheel, Beitrag zur Theorie der Evolutionsstrategie, doctoral thesis, Berlin 1985

[4] J Simkin, C.W.Trowbridge,"Optimization Problems in Electromagnetics", 4th Biennial IEEE Conference on Electromagnetic Field Calculation, Toronto 1990

[5] K.Preis, C.Magele, O.Biro, "Fem and Evolution Strategies in the Optimal Design of Electromagnetic Devices", IEEE Trans. on. Mag., MAG-26, 2181-2183, 1990

[6] K.Preis et al.,"Comparison of Different Optimization Strategies in the Design of Electromagnetic Devices", 4th Biennial IEEE Conference on Electromagnetic Field Calculation, Toronto 1990

[7] N.E.Collins et al, 'Simulated Annealing - An Annoted Bibliography' American Journal of Mathematical and Management Sciences, Vol 8, pp 205 ff, 1988

ELECTROMAGNETIC SCATTERING USING MIXED FINITE AND INFINITE ELEMENTS IN 2D

MALCOLM S. TOWERS, JOHN A. R. MACNAB, and ANDREW McCOWEN

Department of Electrical and Electronic Engineering
University College of Swansea
Swansea U.K.

Abstract - A formulation of Helmholtz's equation has been devolved to allow analysis of electromagnetic backscatter in 2D. The open regions of the problem domain are handled through the inclusion of infinite quadratic wave envelope elements which serve to enforce appropriate far-field attenuation. The remaining regions including the scattering object are modelled using triangular finite elements with standard area shape functions.

Tests have been made to verify the technique on homogeneous dielectric and conducting bodies with known analytic solutions and excellent agreement has been achieved. The method is expected to be particularly useful for inhomogeneous and geometrically irregular problems.

I. INTRODUCTION

Many applications are foreseen for a numerical computation scheme that will efficiently and accurately solve for electromagnetic incidence on arbitrary bodies or scatterers of resonant size or greater. Such applications include the radar signatures associated with targets, electromagnetic compatibility (EMC), the design and siting of antennas and the hazardous effects associated with electromagnetic radiation.

In this paper we discuss the use of a mixed finite/infinite element scheme to solve the time independent form of Maxwell's differential equations. The resulting linear equation system is highly sparce and this fact is used to advantage in their storage and solution. A De Launey mesh is used to discretise the scatterer and the immediate surrounding medium and infinite elements [1] are used to represent the remaining unbounded medium. Initial results in 2D, from the newly developed code SEMS, will be presented for a range of scatterers of resonant dimensions. Far-field scattering is determined directly from the nodal values of the infinite elements and shows excellent agreement with the available analytic solutions. Meshing requirements of the scatterers and the importance of the positioning of the infinite elements will also be discussed.

II. THE FINITE/INFINITE ELEMENT FORMULATION

For a linear source free medium and harmonic time dependance with radian frequency ω each component of the electric and magnetic field vectors satisfies the equations

$$\nabla\times\left(\frac{1}{\mu_r}\nabla\times E\right) - k^2\epsilon_r E = 0 \qquad (1)$$

$$\nabla\times\left(\frac{1}{\epsilon_r}\nabla\times H\right) - k^2\mu_r H = 0 \qquad (2)$$

where E and H are electric and magnetic fields and where ϵ_r and μ_r are relative permittivity and permeability respectively and $k = \omega\sqrt{\epsilon_0\mu_0}$ is the free space wave number. Applying a Petrov-Galerkin formulation to either equation leads to a sparce linear system of equations,

$$Ku = f \qquad (3)$$

where u is a vector of complex field components at nodes of the mesh and right hand side vector f contains boundary integral terms representing the inclusion of incident radiation into the problem [2]. Matrix K is the result of assembling element matrices and applying boundary conditions, for example at the surface of a conductor, and takes the form for

$$K_{ij} = \int\left[\frac{1}{\epsilon_r}(\nabla\times W_i)\cdot(\nabla\times N_j) - \mu_r k^2 W_i N_j\right] d\Omega \qquad (4)$$

where W_i is the global weight function associated with node i, N_j is the global shape function for node j and the integral is taken over the whole mesh. For equation (1), simply interchange ϵ_r and μ_r in equation (4).

In two dimensional problems, such as those reported here, equations (1) and (2) reduce to problems in a single complex field component with TM or TE polarizations respectively and the first term of the integrand in equation (4) reduces to the dot-product of the 2-dimensional gradients of W_i and N_j.

The finite elements radiate from the outer boundary of the finite element mesh into the remaining open space and, following Zienkiewicz et la [3], are provided with shape functions in 2-D of the form

$$N = r^{\frac{1}{2}} M_i(\xi, \eta)\exp\left\{-jk\,(r/R - 1)\right\}. \qquad (5)$$

R is the distance of the element's inner edge from the decay centre chosen for the mesh, r is the distance of point (ξ, η) from the centre and M_i is a standard Lagrange polynomial shape function with quadratic variation in the radial direction, ξ, and linear in η. Reciprocal mapping is used to relate distance r to the element local variable ξ,

$$r = 2R/(1 - \xi) \qquad (6)$$

and clearly r tends to infinity as ξ tends to unity.

If the inverse transformation is applied to M then a polynomal in r^{-1} results. The additional factor of $r^{\frac{1}{2}}$ in equation (5) guarantees the leading term in N decays as $r^{\frac{1}{2}}$. R is normally chosen constant in each element [4] [5] but in the current work R is interpolated through the local variable η. This has the advantage of restoring C_0 continuity to the shape functions but care must be taken to ensure that the integrals in equation (4) are convergent in the infinite elements.

By choosing weight functions as the complex conjugates of the shape functions a wave envelope formulation of the infinite elements results [6] [7] due to cancellation of the complex exponential terms. This allows Gauss-Legendre integration to be used in the standard way. The finite elements used are three noded triangles with standard linear area shape functions. Since these are purely real the resulting matrices coincide with a Galerkin formulation for the finite elements. Employing triangles has the advantage of allowing De Launey meshes to be used with their enormous power and versatility to represent geometrically and materially irregular problems with highly controllable local mesh density.

Far field calculation may be performed directly from the solution on the infinite element nodes, without recourse to elaborate surface integral formulations [8]. Along a radial edge of an infinite element the limit is simply given by

$$\left\{ r^{\frac{1}{2}} u(r) \right\} \longrightarrow R^{\frac{1}{2}}\left(-u_1 + 2\sqrt{2}\,u_2\right) \qquad (7)$$

as $r \longrightarrow \infty$

where u_1 is the field on the inner node, distance R from the decay centre and u_2 is on the outer node, at 2 R from the centre. Linear interplation of these results is used for values between the edges.

III. RESULTS

The problem domain within SEMS is split into 2 regions, see Fig. 1, namely Ω_1 the region meshed with finite elements and Ω_2 the unbounded infinite element region $\sum_2$ is the boundary separating Ω_1 and Ω_2 and $\sum_1$ is one or more closed boundaries within Ω_1 defining the surface of perfect conductors which do not require internal meshing. The initial tests for the code and formulation have been made against the benchmarks of scatterers having an analytic solution

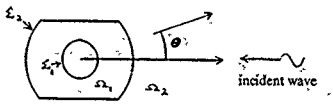


Fig. 1. Definitions of regions and boundaries

The following set of typical results shown in Figs. 2 (a), (b) and (c) give the scattering width for TE incidence on a circular cylinder of $ka = 2.09$ where a is the radius of the cylinder. Fig. 2 (a) is for a perfectly conducting cylinder when Σ_2 is $^1/_8$ wavelength from the surface Σ_1 of the conductor. 61 infinite elements were used to determine the far-field and the corresponding analytic solution was sampled to coincide with the sides of the elements. Reducing the Ω_1 region to just one layer of finite elements slightly deteriorated the solution but by no more than 1 db from the analytic solution Figs. 2 (b) and (c) are for a pure dielectric of permittivity 4 and a lossy one of $4 - j^{10}$ respectively. In both cases the Ω_1 region including the dielectric was meshed with approximately 10 nodes per dielectric wavelength, and only one layer of finite elements outside the cylinder resulting in 47 infinite elements to compute the far-field.

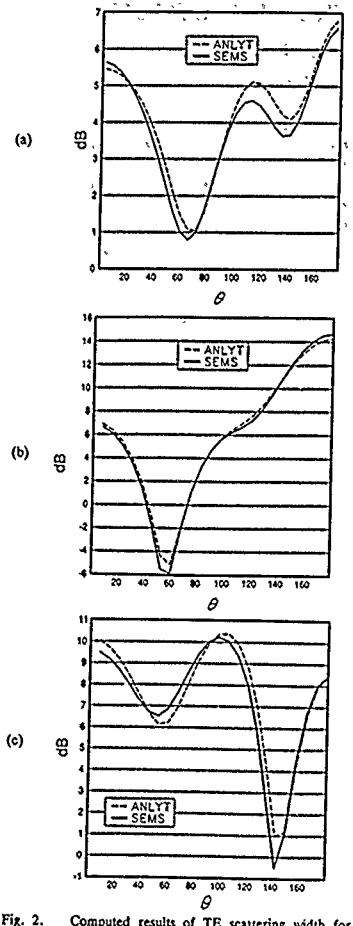


Fig. 2. Computed results of TE scattering width for circular cylinders of radius a with ka = 2.09 and material type (a) perfect conductor, (b) $\epsilon_r = 4$, (c) $\epsilon_r = 4 - j^{10}$.

IV. CONCLUSIONS

A Petrov-Galerkin formulation has been applied to Helmholtz's equations using a mixed finite/infinite element mesh to compute bistatic scattering from 2D objects. Excellent agreement between computed far-field and the analytic solution for a range of circular cylinders has been demonstrated. Post processing has been greatly simplified by computing the far-fields directly from the nodal values of the infinite elements.

Further work including the positioning and shape of the Σ_2 boundary for non-cylindrical scatterers and a comparison of the far-field computed from surface/volume currents with those directly from the infinite elements will also be discussed in the paper.

REFERENCES

1. P. Bettess, "Infinite Elements", Int. J. for Numerical Methods in Engineering, 1977, vol 11, pp 53-64.

2. P. Bettess and O. C. Zienkiewicz, "Diffraction and refraction of surface waves using finite and infinite elements", Int. J. for Numerical Methods in Engineering, 1977, vol 11, pp 1271-1290.

3. O. C. Zienkiewicz, C. Emson and P. Bettess, "A novel boundary infinite element", Int. J. for Numerical Methods in Engineering, 1983, vol 19, pp 393-404.

4. O. C. Zienkiewicz, K. Bando, P. Battess, C. Emson and T. C. Chlan, "Mapped infinite elements for exterior wave problems", Int. J. for Numerical Methods in Engineering, 1985, vol 21, pp 1229-1251.

5. M. J. McDougall and J. P. Webb, "Infinite elements for the analysis of open dielectric waveguides", IEEE Trans. on MT&T, 1989, vol 37, no 11, pp 1724-1731.

6. R. J. Astley, "Wave envelope and infinite elements for acoustical radiation", Int. J. numerical methods fluids, 1983, 3, pp 507-526.

7. P. Bettess, "A simple wave envelope element example", Comms, in Appl. Num. Methods, 1987, vol 3, pp 77-80.

8. E. T. Mayer, "A finite element/boundary integral equation approach to the accurate prediction of radar cross section", 1990, Conf. Proc. 6th Review, Applied Comp.

ACKNOWLEDGEMENT

This work has been carried out with the support of Procurement Executive Ministry of Defence.

The authors wish to acknowledge the work of their colleague G. J. Huang for the development of the 2D De Launay Finite Element Mesh Generator.

NEUTRONS AND DISLOCATIONS IN AN ELECTROMAGNETOELASTIC SOLID

Bogdan Maruszewski

Technical University of Poznań, Institute of Applied Mechanics,
ul. Piotrowo 3, 60-965 Poznań, Poland.

Abstract - The paper deals with a nonconventional thermodynamical modelling of interactions of weak and high energetic neutrons with a defective, deformable semiconducting body. Field equations and boundary conditions are derived. Some examples of interactions are analytically and numerically investigated.

1. INTRODUCTION

Contemporary technologies demand in many situations materials of special properties to obtain expected results which come from simultaneous interactions of physical fields occuring in. The neutrons passing through a crystal occasionally collide with the nuclei of atoms and transfer them kinetic energy. If this recoil energy is large enough the atom concerned will be knocked out of its lattice site and will travel through the crystal, colliding with its neighbours and may be displacing them also from their sites. The end product will be a number of vacant lattice sites and an equal number of displaced atoms. Having produced the interstitials and vacancies the thermal vibration of the crystal lattice causes these so-called point defects to migrate. During migration the point defects may attach themselves forming dislocations. It is well-known that the dislocation line is a reason of an elastic distortion field around it, but the dislocation core, as a singularity of that field, posesses very interesting physical properties. It can be charged for some crystal bonds are broken there, it is a kind of a trap for charge carriers in a semiconductor, it forms a line or a capillary tube which considerably influences the transport of particles through the crystal, it is the line forming a special distribution of spins, and the like. So that, if the neutron beam has the energy sufficiently small the other kinds of interactions with the body occur. Neutrons may diffuse along the dislocation network. Because both have magnetic momemts, there is an interaction between neutrons and electrons.

The proposed theory is applied to several particular cases of interactions between fields mentioned above.

2. THERMODYNAMICAL MODEL

The object of our consideration is a deformable, homogenous, isotropic and defective by dislocations semiconducting body of magnetic properties. It is assumed that the following fields interact with each other:

(i) the *elastic field* described by the stress tensor σ_{ij} and the small strain tensor ε_{ij},

(ii) the *thermal field* described by the temperature T and the heat flux q_i,

(iii) the *electromagnetic field* described by the electromotive intensity $\mathcal{E}_i$ and the magnetic induction B_i,

(iv) the *charge carrier field* described by the concentration n and the flux J^n_i for electrons and the concentration p and the flux J^p_i for holes,

(v) the *neutron field* in the case of a weak energetic beam when its magnetic properties dominate, described by the magnetic moment density μ_i and the flux η_{ij} (the neutron field is understood here as a distribution of travelling and rotating magnetic moments),

(vi) or the *neutron field* in the case of a hard neutron beam described by the particle number density n and the flux b_k,

(vii) the *dislocation field* described by the variable responsible for its density α_{ij} and the flux γ_{ijk}.

Taking into account the relaxation features of the thermal charge carrier, neutron and dislocation fields and basing on the general philosophy of so-called extended irreversible thermodynamics, the independent constitutive variables are represented by the set

$$\mathcal{C} = \Big\{ \varepsilon_{ij}, \mathcal{E}_i, B_i, T, T_{,i}, n, n_{,i}, p, p_{,i}, \mu_i, (\text{or } n)$$
$$\alpha_{ij}, \alpha_{ij,k}, q_i, J^n_i, J^p_i, \eta_{ij}, (\text{or } b_k)\, \gamma_{ijk} \Big\} \qquad (2.1)$$

This specific choice shows that magnetic exchange effects are discarded and that we ignore the relaxation features of the mechanical field (viscosity) so that σ_{ij} is not in the set (2.1). Moreover, we assume that the contribution of the neutron and dislocation fields into processes is of such a character that μ_i and α_{ij} can be understood as the internal variables.

Neglecting mass of neutrons in comparison to the solid we take the continuity equation in its classic form (ρ denotes the density)

$$\dot{\rho} + \rho\, v_{i,i} = 0. \qquad (2.2)$$

The balances of *momentum*, *moment of momentum* and *internal energy* for the considered body read assuming that the electrical polarization is practically negligible (the magnetic properties dominate)

$$\rho \dot{v}_i - [\sigma_{ji,j} + e_{ijk}(J^n_j + J^p_j)B_k + B_{i,k}M_k + f_i] = 0$$
$$e_{ijk}\sigma_{jk} + c_i = 0 \qquad (2.3)$$
$$\rho\dot{e} - [\sigma_{ji}v_{i,j} + (J^n_k + J^p_k)\mathcal{E}_k - M_k\dot{B}_k - q_{k,k} + \rho r] = 0,$$

where v_i is the velocity of the body point, M_k denotes the magnetization, f_i is the body force, c_i is the couple, e is the internal

energy density and r is the heat source. The electromagnetic field is governed by the following *Maxwell equations*:

$$e_{ijk}E_{k,j}+\frac{\partial B_i}{\partial t}=0\,;\qquad D_{i,i}=\rho Z\,,$$

$$e_{ijk}H_{k,j}-(J_i^n+J_i^p)-\rho Z v_i-\frac{\partial D_i}{\partial t}=0\,,$$

$$B_{i,i}=0\,, \tag{2.4}$$

where [11,18] .

$$\mathcal{E}_i=E_i+e_{ijk}v_jB_k\,,\qquad D_i=\varepsilon_oE_i\,,$$

$$Z=n+\bar{n}-n_o+p+\bar{p}-p_o\,, \tag{2.5}$$

$$B_i=\mu_o(H_i+M_i)\,,$$

ε_o,μ_o denote the permittivity and permeability of vacuum, H_i is the magnetic field intensity, $\bar{n}$, $\bar{p}$ are charge densities of fixed impurities and n_o,p_o are the equilibrium concentrations of electrons and holes, respectively. The relation (2.5) should be written as $B_i=\mu_o(H_i+M_i+\mu_i)$ for the weak neutron beam case, but as $|\mu_i|\ll|M_i|$ we have kept the classical expression (2.5)4. According to the evolution of the total charge we are interested only in *the balances of charge carriers* for the fixed charge of impurities is rather weakly engaged into interactions. Their forms read

$$\dot{n}+J^n_{i,i}=g^n-r^n$$
$$\dot{p}+J^p_{i,i}=g^p-r^p\,, \tag{2.6}$$

with *the evolution equations of electron and hole fluxes*

$$\overset{*}{J}{}^n_i=\mathcal{J}^n_i(C)$$
$$\overset{*}{J}{}^p_i=\mathcal{J}^p_i(C)\,. \tag{2.7}$$

where g^r, g^p, r^n and r^p denote the generation and recombination of the electrons and holes, respectively. Since the dislocations during interactions as well as the neutrons can move through a crystal the evolution equations for the above fields take the forms

$$\overset{*}{\mu}_i+\eta_{ij,j}-\mathcal{M}_i(C)=0\qquad \dot{n}+b_{k,k}-\mathcal{N}(C)=0$$
$$\overset{*}{\eta}_{ij}-\mathcal{H}_{ij}(C)=0\qquad \overset{*}{b}_k-\mathcal{B}_k(C)=0$$
$$\overset{*}{\alpha}_{ij}+\gamma_{ijk,k}-\mathcal{A}_{ij}(C)=0 \tag{2.8}$$
$$\overset{*}{\gamma}_{ijk}-\mathcal{V}_{ijk}(C)=0\,,$$

The last law concerns *the heat flux evolution equation*

$$\overset{*}{q}_i=Q_i(C)\,. \tag{2.9}$$

The superimposed dot and asterisk denote the material and Zaremba-Jaumann time derivatives, respectively.

We must comment the equation (2.8). In a case of high energetic neutron beam $\mathcal{N}(C)$ is practically negligible if nuclear reactions do not occur within cascade of collisions. Otherwise $\mathcal{N}(C)$ does not vanish and $\mathcal{M}_i(C)$ form a source of additional spins coming from the produced neutrons. Then the relaxation time of the neutron field for the hard beam tends to zero because the beam is very stable and inertial on influences from outside. This results in $\mathcal{B}_k(C)=0$ and means that practically only transport features of that field should be considered. According to the dislocation field the term $\mathcal{A}_{ij}(C)$ is responsible for a production of dislocations coming from the interaction of the hard neutron beam with the lattice. If we, however, deal with the weak energetic neutrons the magnetic interactions dominate in processes, $\mathcal{M}_i(C)$ describes couplings with the remaining fields in the body and the relaxation time of the neutron field is relatively long. According to $\mathcal{A}_{ij}(C)$ it is practically negligible for the weak energetic neutron beam interaction as a source of dislocations and can occur in the case of elastic deformation as well as for couplings with such physical fields as the electromagnetic, thermal, and the like.

All the fundamental laws (2.2)-(2.8) should not contradict the second law of thermodynamics. Thus all the considered physical processes have to be restricted by the following entropy inequality

$$\dot{\rho s}+\Phi_{i,i}-\frac{\rho r}{T}\geq 0 \tag{2.10}$$

where s denotes the entropy density and Φ_i is the entropy flux.

The most difficult step in constructing the theory of interactions is its constitutive part. On introducing the set

$$\mathcal{Z}=\{\sigma_{ij},M_k,c_i,e,r,g^n,g^p,r^n,r^p,\mathcal{M}_i\ (\text{or }\mathcal{N}),\mathcal{H}_{ij}(\text{or }\mathcal{B}_k)$$
$$J^n_i,J^p_i,Q_i,\mathcal{A}_{ij},\mathcal{V}_{ijk},s\,,\Phi_i\} \tag{2.11}$$

we look for the constitutive relations in the form

$$\mathcal{Z}=\mathcal{Z}(C)\,. \tag{2.12}$$

On analyzing the entropy inequality together with (2.2)-(2.7) with the help of Liu's theorem we use the free energy and flux-potential function

$$\Psi=e-Ts\,,\qquad K_i=\rho v_i\Psi-T\Phi_i \tag{2.13}$$

as the most proper ones for the considered processes.

As a result we obtain general definitions of the constitutive relations (laws of state) basing on the free energy density, the fluxlike relations leading to the explicit form of the entropy flux basing on the flux-potential function and a residual inequality which the kinetic constitutive relations can be defined from Some particular examples of interactions are analytically and numerically investigated.

DYNAMIC BEHAVIOR OF THIN SHELL STRUCTURE UNDER MOVING / TRANSIENT MAGNETIC FIELD

TOSHIYUKI TAKAGI AND JUNJI TANI
The Institute of Fluid Science, Tohoku University
Katahira 2-1-1, Aoba-ku, Sendai 980, Japan

Abstract-This paper describes the analyses of the coupled electromagnetic-mechanical problems on the dynamic behaviors of a thin plate and a shallow arch under moving / transient magnetic field. Numerical analyses, considering the coupling effect between magnetic field and defection, are performed for the evaluation of eddy current, magnetic field, force and vibration.

I. INTRODUCTION

Thin electrically conducting plate and shallow arch are used in high magnetic field machines. They receive transient and/or impulsive electromagnetic forces, deflect dynamically and sometimes show magnetoelastically unstable behaviors.

The coupling effect of a beam has been investigated by L. R Turner [1], T. Morisue [2], S. Matsuda [3] and so on. They treated the rigid-body movement and the elastic beam deflection under uniform time-changing field. Dynamic responses of shallow arches under mechanical forces have been already studied [4]. Humphreys derived the nonlinear equation of motion for an arch under an impulsive mechanical load considering an axial force [4]. We have already shown the preliminary results for the dynamic behavior of a shallow arch under transient electromagnetic force [5].

In this paper, we show the numerical calculation procedure for the coupling effect between electromagnetic field and vibration for both a thin plate and a shallow arch.

II. NUMERICAL ANALYSES METHOD

A. Eddy Current Analysis

A thin elastic isotopic homogeneous shell with electrical conductivity σ and magnetic permeability μ_0 is set under a coil field.

From the Ohm's law, the following equation is obtained,

$$J=\sigma(E+u\times B) \tag{1}$$

where J is current density and u is deflection velocity vector.

Since $\mathrm{div}\, J=0$, the current vector potential T is defined as, $J=\mathrm{rot}\, T$. Using the Faraday's law and eq (1), we get the following equation.

$$\mathrm{rot}(\mathrm{rot}\, T)=-\sigma\frac{\partial B}{\partial t}+\sigma\mathrm{rot}(u\times B) \tag{2}$$

We divide the magnetic induction into external magnetic induction B_0 and that induced by eddy current B_e. We finally obtain the governing equation using the current vector potential for a thin plate considering the coupling effect as,

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial T^2}{\partial y^2}=\sigma\frac{\partial}{\partial t}(B_{0z}+B_{ez})-\sigma V_b \tag{3}$$

$$V_b=B_{0x}\frac{\partial^2\delta}{\partial x\partial t}+B_{0y}\frac{\partial^2\delta}{\partial y\partial t}-\frac{\partial B_{0z}}{\partial z}\frac{\partial\delta}{\partial t} \tag{4}$$

where δ is the lateral displacement.

B. Deflection Analysis

B.1 Plate deflection

The governing equation for the deflection of the thin plate considering the Lorentz force is expressed as follows:

$$D\left(\frac{\partial^4\delta}{\partial x^4}+2\frac{\partial^4\delta}{\partial x^2\partial y^2}+\frac{\partial^4\delta}{\partial y^4}\right)+\rho h\frac{\partial^2\delta}{\partial t^2}=hB_y j_x-hB_x j_y \tag{5}$$

where D, h and ρ are bending rigidity, plate thickness and mass density respectively. Coupling with eddy current calculation, we calculate the plate deflection based on the finite element method with the nonconforming triangular elements using the modal analysis [6]

B.2 Arch deflection

The equation of motion for a shallow arch was given by Humphreys when a mechanical impulsive force acts [4]. By adding the structural damping term to the Humphreys' equation the following non-linear differential equation can be written,

$$EI\frac{\partial^4\delta}{\partial x^4}-S\frac{\partial^2}{\partial x^2}(y_0-\delta)+\rho A\frac{\partial^2\delta}{\partial t^2}+C\frac{\partial\delta}{\partial t}=Wq \tag{6}$$

where EI, S, y_0, ρ, A, C, W and q are flexural rigidity, resultant axial force, initial arch shape, mass density, cross-section, damping coefficient, arch width, and Lorentz force respectively.

Using six vibration modes, we obtain the nonlinear equations of motion and solve these equations iteratively using Newmark β method coupled with eddy current calculation.

III. RESULTS AND DISCUSSION

In this study we performed two numerical calculations.

(1) Dynamic behavior analysis of a shallow arch under a pulse coil field [7].
(2) Dynamic behavior analysis of a thin plate under moving magnetic field [8].

A. Shallow Arch

An arch test piece is set with both ends clamped as shown in Fig.1. The length, the width, the thickness, the arch radius of an aluminium test piece are 100, 20, 0.2 and 500mm, respectively. Fig.2 shows the measured dynamic responses of the arch when maximum pulse coil current was 393A (pulse width, about 4msec). The numerical results for the response with and without considering the coupling effect are shown in Fig.3. Vibration behavior (such as amplitude, frequency, vibration mode etc.) with the coupling effect almost agreed with that of Fig.2.

Fig.4 shows the relation between the coil current and the maximum deflection. Both numerical and experimental results might show a kind of snap-through buckling behavior. We define the critical current when the maximum deflection becomes a half of the arch height (H). As shown in the figure the numerically predicted critical current with the coupling effect was within the scatter of experimental ones.

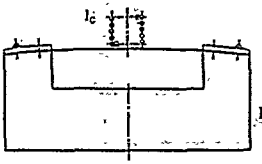
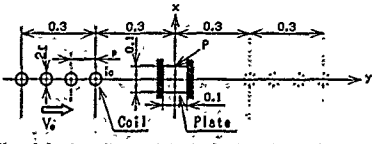


Fig.1 Test apparatus



r: radius of circular coils p: pitch of coils V_c: velocity of a series of coils
I_c: current of coils r=0 025, p=4r

Fig 5 - Analytical model

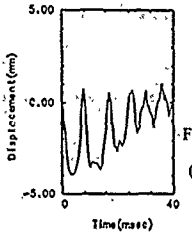
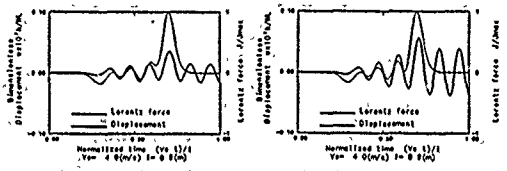
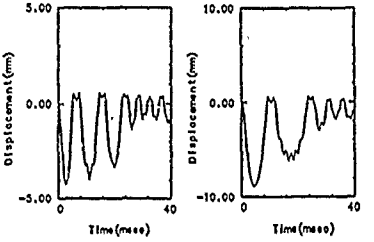


Fig 2 Dynamic response of an arch (Experiment, I_c=393A)



(a) With coupling terms (b) Without coupling terms

Fig 6 Displacement and Lorentz force, I_c=0.5(MA)



(a) With coupling term (b) Without coupling term

Fig.3 Dynamic response of an arch (Analysis, I_c=393A)

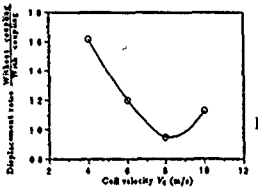


Fig.7 Coupling effect

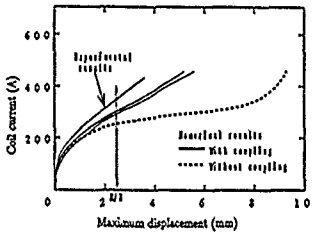


Fig.4 Pulse coil current vs maximum deflection

B. Thin Plate

A thin elastic isotropic homogeneous copper square plate was set under moving magnetic field as shown in Fig 5. The width and the length of the plate are both 0 1(m). The thickness of the plate is 0 1(mm). It was fixed at two sides and free at other sides. A series of four moving circular coils on x-axis generates moving field. The coils move from the positions of solid circles to those of broken circles. Figs 6(a) and 6(b) shows the dynamic response of the plate in the case of the coil current of I_c=0.5(MA) and the coil velocity of V_c=4m/sec. As shown in the figures, when we considered the coupling effect, the amplitude of vibration became smaller because of magnetic damping effect Fig.7 shows the relation between the ratio of maximum displacement with the coupling effect over that without the effect and coil velocity. The ratio became minimum (0 95) at V_c=8(m/s) and it was less than one. This means that the magnetic damping coefficient changes as a function of the coil velocity and a negative damping effect exists at a certain velocity.

IV. CONCLUSIONS

We showed a numerical analysis method of dynamic responses of thin shell structure under magnetic field. The method was applied to the dynamic behavior analysis of a shallow arch under a pulse coil field and a thin plate under moving magnetic field. In both cases we showed the differences between numerical results with and without the coupling effect. Therefore we can conclude that the coupling calculation is quite important to evaluate the dynamic response of thin shell structures under magnetic field.

We wish to acknowledge Mr. K. Ohtomo, Mr. S. Matsuda and Mr. Y. Mohri for doing experiment and numerical calculation.

REFERENCES

[1] L. R. Turner and T Q. Hua, Electromagnetomech. Interactions in Deformable Solids and Struct., (1987), pp 81-86.
[2] T. Morisue, IEEE Trans. on Mag., Vol 26(1990), pp.540-543.
[3] S. Matsuda et al., Proc. 4th Int. IGTE Symp. and Europ. TEAM Workshop, Graz, Austria(1990), pp.259-269.
[4] J. S. Humphreys, AIAA J. Vol.4(1966), pp.878-886.
[5] T. Takagi et al., Proc. 4th Int. IGTE Symp. and Europ. TEAM Workshop, Graz, Austria (1990), pp.107-112.
[6] T Takagi et al., Intern. J. Appl. Electromagn. Mater. Vol.1 (1990), pp.147-154.
[7] T. Takagi et al, 2nd Japanese-Polish Joint Seminar on Electromagn. Phenom. Mater Ona, Japan (1991) (Submitted).
[8] T. Takagi et al., Intern. Symp. on Appl. of Electromagn Forces, Sendai, Japan (1991) (Submitted).

BUCKLING BEHAVIORS OF CYLINDRICAL SHELLS DUE TO IMPULSIVE ELECTROMAGNETIC FORCES

[1]K. Iŋki, [2]Y. Kawamoto, [2]M. Nemoto, [3]M. Hashimoto,
[4]M. Tsuchimoto, [2]M. Kushiyama and [4]K. Miya

[1]Mitsubisi Atomic Power Industries, Inc., 2-4-1, Shibakouen, Minato, Tokyo, Japan.
[2]Mitsubisi Heavy Industries, Ltd., 1-1 Akunoura-machi, Nagasaki, Japan.
[3]University of Industrial Technology, Sagamihara, Kanagawa, Japan.
[4]Nuclear Engineering Research Lab., University of Tokyo, Tokai-mura, Ibaraki, Japan.

Abstract- The buckling characteristics of a cylindrical shell are experimentally studied to determine the effects of impulsive magnetic forces, and compared to calculation results. Dynamic effects including the electromagneto-mechanical coupling are observed at impulsive loads of high frequency 1-10kHz.

I. INTRODUCTION

Large impulsive magnetic forces act on the vacuum vessel and other components of fusion devices during plasma disruptions. Therefore, the buckling problem is one of the important considerations during the design period of such devices. However, the buckling behaviors of cylindrical or toroidal shells under impulsive magnetic forces is not well known. As the first step towards addressing this problem, experiments and analyses have been done on cylindrical shells.

II. EXPERIMENTAL DEVICE

A capacitor-bank and a working solenoidal coil are the main elements of the experimental apparatus as shown in Fig.1 [1]. Ignitron switch controllers are prepared to make impulsive sinusoidal half exciting current waves as shown in Fig.2. The working coil is 100mm in length and 60mm in diameter. Two kinds of capacitor-bank are used: C=1400μF (200μFx7) and 18μF (6μFx3). The coil current peak time was changed from Δt=0.025mS(10kHz) to 0.2mS(1.25kHz) by adjusting L and C.

The test specimens are cylindrical shells, 53mm in outer diameter and 100mm and 60mm in length with a variety of thickness: h=1, 0.5 and 0.14mm. The cylindrical shell is located inside the coil. Aluminum alloys are selected as test materials because of their high electric conductivity and low density, which are preferable for testing deformation and buckling behaviors of the shell. Two aluminum alloys, A1070-O and A5052-H14, are used to investigate the yield strength dependence on buckling behaviors. The 0.2% yield strengths are 40MPa and 220MPa, respectively. The eight items are measured as shown in Fig.1, avoiding electromagnetically induced noise and responding to high speed phenomena.

III. MEASUREMENT AND ANALYSIS OF EDDY CURRENTS

An induced one-turn current is measured by a differential pick-up method with a couple of small coils that are placed on both sides of the cylindrical shell as shown in Fig.1. However, this measurement can not be applied to the experiment with large deformation of the shell. The experimental data of an induced one-turn current are compared to calculated results as shown in Fig.2. They agree well with each other as long as the deformation is small.

Eddy currents induced on the cylindrical shell have been calculated with MATEX [2] and EDDYCUFF [3], that uses a finite element circuit method (the network mesh method with finite elements) with the normal component of the current vector potential Tn. The maximum induced one-turn current is ~70% of the maximum coil ampere-turns in the case of f=1.25kHz, h=1mm, ρ=2.9μΩcm as shown in Fig.2. The calculated results are summarized for the maximum induced one-turn current in Fig.3. The electromagneto-mechanical coupling is not considered in these results.

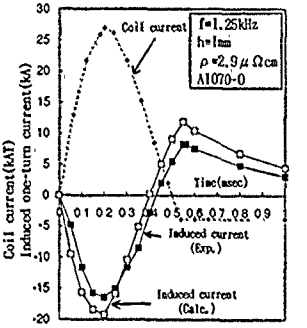


Fig.2 Comparison of measured and calculated one-turn currents induced on the cylindrical shell.

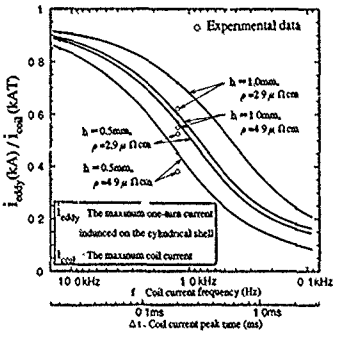


Fig.3 Calculated ratio of induced one-turn current to coil current as a function of coil current frequncy.

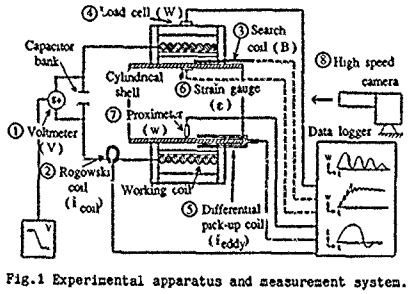


Fig.1 Experimental apparatus and measurement system.

IV. DEFORMATION AND BUCKLING.

Transient deformation during the loading test of an impulsive magnetic force was observed by a high speed camera. The pictures in Photo 1, taken every 100μs, show the deformed shapes of the cylindrical shells. The buckling wave number in the circumferential direction seems to increase as the load duration becomes short, such as a typical wave number n=6~8 at the 1.25~10kHz dynamic loads.

The residual deformation of the cylindrical shells was investigated after the loading test. An example of buckled cylindrical shells is shown in Photo 2. Figures 4 and 5 show the buckling loads experimentally obtained in the tests of the A1070-O and A5052-H14 shells, respectively. The buckling loads are shown as a function of the radius to thickness ratio (a/h) of the cylindrical shell, with the parameter of impulsive current frequency. They are compared to calculated results of elastic, plastic and elastic-plastic buckling for static loads. Elastic-plastic buckling analysis has been carried out using BOSOR [4]. Buckling of A1070-O shells occurs in the plastic range as shown in Fig.4. On the other hand, buckling of A5052-H14 shells occurs in the elastic range. Experimental results are higher than estimated static buckling loads at high frequency.

V. DISCUSSION

The difference between the experimental and calculated results is explained by combined dynamic effects. Structure response analysis for dynamic loads has been performed and the dynamic response factor is calculated to be 1.04~1.56 at the 1.25~10kHz impulsive loads. Figure 6 and 7 shows the dynamic effects of the buckling load as a function of the peak load time. The dynamic buckling load can be predicated by the elastic-plastic buckling load divided by the dynamic response factor as shown in Fig.6 and 7. In the results, the large difference between the experimental data and the predicated dynamic buckling loads at high frequency is thought to be due to the dynamic effects of electromagneto-mechanical coupling, buckling strength and material properties. It is reported that buckling strength and material properties such as yield strength change at short-duration impulsive loads.

The effect of electromagneto-mechanical coupling can be calculated from deformation data derived from high speed pictures, such in Photo 1. It is estimated that the magnetic pressure on the cylindrical shell can be reduced 10~50% by the coupling effect in typical buckling conditions.

VI. CONCLUSION

Buckling characteristics of a cylindrical shell are experimentally obtained for impulsive electromagnetic loads of 10kHz and 1.25kHz. Buckling loads experimentally observed are much higher than those expected from static buckling analysis and dynamic structure response analysis at impulsive loads of high frequency. This difference can be explained by the electromagneto-mechanical coupling and dynamic effects of buckling strength and material properties.

Photo 1. Transient deformation of the cylindrical shell cross-section. Time step:δt=100μs (A5052-H14, f=1.25kHz, h=0.5mm, Eo=1400V, i_{coil}=220kAT).

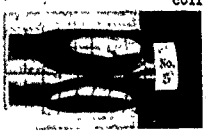


Photo 2. Appearance of the buckled cylindrical shell.

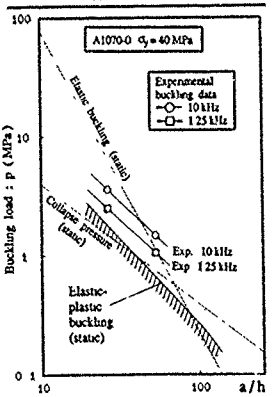


Fig.4 Impulsive electromagnetic buckling characteristics (A1070-O).

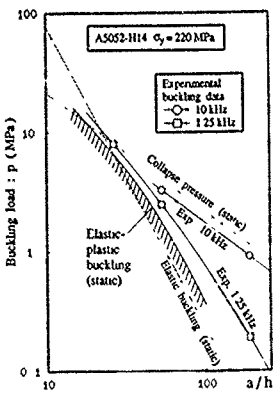


Fig.5 Impulsive electromagnetic buckling characteristics (A5052-H14).

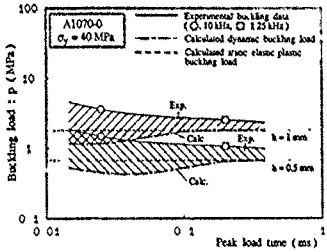


Fig.6 Dynamic effects of impulsive electromagnetic buckling as a function of peak load time (A1070-O).

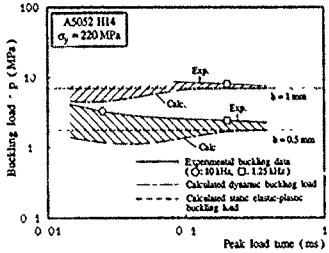


Fig.7 Dynamic effects of impulsive electromagnetic buckling as a function of peak load time (A5052-H14).

REFERENCES

[1]M.Nemoto et al., Proc. of ISEM-Sendai (1991), to be published.

[2]K.Ioki et al., Nuclear Engineering and Design/Fusion 3 (1985) 63.

[3]A.Kameari, Journal of Comp. Physics, 49 (1982) 389.

[4]D.Bushnell, Computers and Structures 6 (1976) 221.

MAGNETO-ELASTIC BUCKLING OF SUPERCONDUCTING COILS

A.A F. VAN DE VEN
Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract Systems of superconducting coils can buckle whenever the electric current through the coils exceeds a certain critical value. A variational method which uses as admissible magnetic fields the so-called Biot-Savart fields is presented. An evaluation for slender systems of curved beams is given. The method is applied to a conductor in the shape of a flat spiral coil.

INTRODUCTION

A superconducting body in the shape of a slender curved beam is placed in a vacuum and carries a prescribed current I_0. The own field of the conductor, generated by I_0, causes large Lorentz forces between the seperate parts of the conductor. Due to this the system can become mechanically unstable and buckle whenever I_0 becomes too large ($I_0 > I_{0cr}$). In a series of earlier papers, [1] - [3], the present author together with his co-workers presented two methods for the calculation of the buckling current I_{0cr}. The first one is a variational method, the second one a more direct one starting from a formula for the Lorentz force on a conducting slender coil that is derived from the law of Biot and Savart in a way as shown in [4], Sec. 2.6. Both methods have been applied to calculate I_{0cr} for systems of parallel rods and rings ([2] - [3]). Comparison of the results shows that under certain (moderate) restrictions the agreement between the two methods is good. Since exact fields are needed for the variational method as employed in our earlier papers this is a mathematically exact but also very complicated method. On the other hand the Biot-Savart method is less precise but easier in use. An approach that combines the advantages of both methods is presented here. This combined method is a variational one. However, in this formulation an admissible set of fields is chosen on the basis of the law of Biot and Savart. Since the differences between our two previous methods were so small it seems logical to expect that the combined method will also yield good results. Some of our results indeed show that the correspondence becomes even better when using the combined method. The combined method becomes extremely simple when applied to slender spiral or helical coils. An example of such a calculation will be given, but first we shall present the main lines of the variational method and show how this leads to the combined method.

MAIN LINES OF VARIATIONAL METHOD

The fundamental quantity in the variational method as presented in [1] is the functional J, which is equal to

$$J = W - I_0^2 K, \tag{1}$$

where W is the elastic energy of the deformed (buckled) body and (μ_0 is the permeability in vacuum; we use summation convention and $_{,i} = \partial/\partial x_i$)

$$I_0^2 K = -\frac{1}{2\mu_0}\int_{\partial G}[(\psi + B_k U_k)_{,j}(B_j U_i N_i + \psi N_j)]dS. \tag{2}$$

according to [3], Eq. (2.22) (here, we have already neglected some terms that become small for slender systems). In (2), ∂G represents the boundary of the conducting body and N is the unit normal on it. Furthermore, U is the displacement in buckling of the conductor, B is the rigid-body magnetic field in the vacuum G^+ and ψ is the perturbed (due to U) magnetic potential in G^+. The displacement U must be chosen in such a way that it is characteristic for the deflection of the specific slender body under consideration and for the type of buckling that is assumed to take place. The field B and the potential ψ have to satisfy the constraints (fields that satisfy these constraints are called admissible)

i) $\quad \operatorname{curl} \mathbf{B} = \mathbf{O}\ ,\ x \in G,\ \ \mathbf{B} \to \mathbf{O}\ ,\ |x| \to \infty,$

$$\text{(and Ampère's law)}\ \int_C (\mathbf{B}, \mathbf{t})ds = \mu_0 I_0\ ,\ C \in G^+; \tag{3}$$

ii) $\quad \partial\psi = 0;\ x \in G^+\ ;\ \psi \to 0\ ,\ |x| \to \infty. \tag{4}$

Our variational principle as presented in e.g. [3] states that the first variation of J with respect to admissible U, B and ψ must be zero. On this basis we proceed as follows. we first choose admissible fields B and ψ and thereafter we take the first variation of J with respect to U equal to zero. This leads us to an eigenvalue problem from which I_{0cr} can be calculated.

ADMISSIBLE FIELDS FOR SETS OF PARALLEL STRAIGHT CURRENT CARRIERS

Consider one infinitely long straight current carrier (with e_3 along the axis and $x = xe_1 + ye_2 + ze_3$). Let R be a characteristic measure of length for the cross-section of the rod and suppose that the rod is periodically supported over lengths l. The rod is called slender if $0 < \delta := R/l \ll 1$. We only consider slender rods and we neglect terms that are $0(\delta^3)$ with respect to unity. Finally, we assume that the rod buckles in the e_1 direction, so that $\mathbf{U} = U(x)\mathbf{e}_1$

Application of the law of Biot and Savart to the undeformed (for B) and to the linearly deformed rod (for ψ) yields the following two admissible fields

$$\mathbf{B}(x) = -\frac{\mu_0 I_0}{2\pi}\frac{(ye_1 - xe_2)}{(x^2 + y^2)}, \tag{5}$$

$$\psi(x) = \frac{\mu_0 I_0}{4\pi}\frac{2y}{(x^2 + y^2)}U(z)\left[1 + 0\left(\delta^3\right)\right]. \tag{6}$$

It can be easily checked that B and ψ satisfy the constraints (3) and (4), however ψ only up to $0(\delta^3)$-terms.

For a set of n parallel rods the admissible fields can be found from (5) and (6) by simple superposition. Using these fields in (2) we obtain for a set of n equidistant rods (distance between two rods is $2a$, $a \ll l$)

$$I_0^2 K = \frac{\mu_0 I_0^2 \kappa}{32\pi a^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\frac{1}{(j-i)^2}\int_{-l}^{l}\left[U^{(i)}(z) - U^{(j)}(z)\right]^2 dz\ , \tag{7}$$

where $U^{(i)}$ is the lateral deflection in the plane of the rods (i.e. in the e_1-direction) of the i^{th} rod, and

$$\kappa = \kappa(m) = \frac{1 - 4m^2 + \frac{3}{2}m^4}{(1-m^2)^{\frac{3}{2}}} = 1 + 0(m^2), 0 < m := \frac{R}{2a} < \frac{1}{2}\ , \tag{8}$$

for a circular cross-section, radius R.

For W we can use the familiar expression for the elastic bending energy of a slender beam. For $n = 2$ this yields a value for I_{0cr} which is just in between the values from the Biot-Savart method (for which $\kappa = 1$) and the exact variational method (cf [2])

SLENDER CURVED BEAMS

Let us consider a superconducting coil in the shape of a slender curved beam like a (flat) spiral or a (straight or toroidal) helix. We call the total structure slender if the distance $2a$ between two branches is much smaller then some global measure of length b of the structure (e.g. a radius of curvature of the central line or the total length of the coil). Take two distinct branches (or turns) of the spiral or helix, let P_1 be a point on one branche and P_2 that point on the other branche that has the shortest distance to P_1. For the structures we consider here, the tangent lines on the central lines in P_1 and P_2 are (nearly) parallel. For slender systems, i.e. under the neglect of terms of $0(a/b)$ with respect to unity, this has the following consequence: The contribution to the K-integral from a line-element in P_1 and due to the interaction with the other branche through P_2 is equal to the one found for the case of two straight parallel rods through P_1 and P_2 and tangent on the curve.

This statement was already checked in [3] for systems of two rings and there proven to be true for small values of (a/b), where b is the radius of the ring. With this we now have the disposal of an easy algorithm to calculate the K-integral for such complex systems as a helical or spiral conductor.

SPIRAL CONDUCTOR

A flat spiral of n turns is given by the following relation for its central line

$$b(\varphi) = R_0 + h\varphi\ ,\quad 0 \le \varphi \le 2\pi n\ . \tag{9}$$

The distance between two adjacent turns is

$$2a = b(\varphi + 2\pi) - b(\varphi) = 2\pi h\ . \tag{10}$$

The system is called slender if

$$\forall_{\varphi\in[0,2\pi n]}\quad \frac{a}{b(\varphi)} = \frac{\pi h}{b(\varphi)} \ll 1\ . \tag{11}$$

When relevant, we take the cross section circular, radius $R (R < a)$.

We assume that the coil buckles in its plane and that the pertinent displacement is given by (in polar coordinates)

$$\mathbf{U} = \mathbf{U}(\varphi) = U(\varphi)\mathbf{e}_r + V(\varphi)\mathbf{e}_\varphi\ . \tag{12}$$

The spiral is taken inextensible, yielding

$$U(\varphi) + V'(\varphi) = 0\quad ,\quad (' = \frac{d}{d\varphi})\ . \tag{13}$$

The elastic energy due to in plane bending is then (EI is the bending stiffness)

$$W = \frac{1}{2}EI\int_0^{2\pi n}\frac{1}{b^3(\varphi)}[U''(\varphi) + U(\varphi)]^2 d\varphi\ . \tag{14}$$

A value for the K-integral can be obtained from the assumed analogy with formula (7). To this end we introduce for $\varphi \in \mathcal{L}_i, i \in [1, n]$, where $\mathcal{L}_i$ stands for the i^{th} turn of the spiral, the pole angle θ by

$$\varphi = 2\pi(i-1) + \theta\quad ,\quad 0 \le \theta < 2\pi\ , \tag{15}$$

and the displacement of $\mathcal{L}_i$ by

$$U(\varphi) = U^{(i)}(\theta) = U(\theta + 2\pi(i-1))\ , \tag{16}$$

and, analogously,

$$b(\varphi) = b_i(\theta) = b(\theta + 2\pi(i-1)) = \\ = R_0\left\{1 + \frac{h}{R_0}[2\pi(i-1) + \theta]\right\}\ . \tag{17}$$

Replacing in (7). $dz \to b_i(\theta)d\theta$ and $a \to \pi h$ we thus obtain

$$I_0^2 K = \frac{\mu_0 I_0^2 \kappa}{32\pi^3 h^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{(j-i)^2}\int_0^{2\pi}[U^{(i)} - U^{(j)}]^2 b_i d\theta\ , \tag{18}$$

where the integral must be taken with respect to $\mathcal{L}_i$.

To obtain a numerical value for I_{0cr} we have followed the following procedure.

i) We have discretized the displacement $U(\varphi)$ in $N+1$ points, the nodal displacements being U_k , $k \in [0, N]$, with $U_0 = U_N = 0$ (the spiral coil is simply supported in begin and end points).

ii) We have calculated the first variation δJ of J with respect to U_k and put this equal to zero, yielding a linear eigenvalue problem for U_k of the form

$$A_{kl}(I_0)U_l = 0\quad ,\quad k,l \in [1, N-1]\ . \tag{19}$$

iii) The lowest eigenvalue I_0 (following from $\det A = 0$) is the looked for buckling current I_0.

We worked out this scheme for a spiral coil for which $h/R_0 = 0.01$ and for some values of n. The results are shown in the Table below in which

$$\lambda := \frac{\mu_0 I_0^2 \kappa R_0^4}{48\pi^3 h^2 EI}\ . \tag{20}$$

More numerical results will be published in the forthcoming paper [5].

TABLE 1. Normalized buckling currents for $h/R_0 = 0.01$ and for some numbers of coils.

n	2	3	4	5
$\lambda * 10^3$	16.30	8.40	5.19	3.50

ACKNOWLEDGEMENT. The author wishes to thank L.G.F.C. van Bree for his assistance in the numerical calculations.

REFERENCES

[1] Lieshout, P.H. van, Rongen, P.M.J., and Ven, A.A.F. van de, J. Eng. Math. 21 (1987) 227.

[2] Lieshout, P.H. van, Rongen, P.M.J., and Ven, A.A.F. van de, J. Eng. Math. 22 (1988), 143.

[3] Smits, P.R.J.M , Lieshout, P.H. van, and Ven, A.A.F. van de, J. Eng. Math. 23 (1989), 157.

[4] Moon, F.C , Magneto-solid mechanics (Wiley, New York, 1984).

[5] Ven, A.A.F. van de, and Bree, L.G.F.C. van, Buckling of superconductive structures; a variational approach using the law of Biot and Savart, forthcoming.

Coupling of Transient Fields, Circuits, and Motion Using Finite Element Analysis

Sheppard J. Salon
Rensselaer Polytechnic Institute, Troy, New York USA

Abstract

The transient analysis of a coupled electro-mechanical system is undertaken. The system consists partly of spatial regions, which may support magnetic fields, that are modelled by finite elements The regions may be attached to external electrical sources and circuits, and may also be capable of rigid body motion with respect to one another A method for coupling the electric circuit transient equations, transient magnetic field finite element equations, and the transient mechanical motion equations is described. Only the external source variation is assumed to be known; all other field, circuit, and mechanical motion quantities are treated as unknowns and calculated. Equations for transient analysis of a general, 2-dimensional, planar, non-linear, voltage-excited system are derived in detail. The Galerkin formulation, time-discretization, and linearization of these equations are presented. The resulting global system of coupled electro-mechanical equations is assembled and investigated Included are examples of the application of the proposed method to perform transient analysis of practical coupled electro-mechanical systems.

1 Introduction

The standard procedure of using finite element analysis to approximate magnetic field quantities within a fixed device or region is well known. The user describes the problem geometry and material characteristics, sets the boundary conditions, and specifies numerically all current densities, which act as the source of the magnetic field. The region of interest is then discretized in space into a mesh, and the finite element field approximation equations are set up and solved. The solution consists of a set of approximations for the field potential at each node of the mesh.

Such a procedure is inadequate for a large class of practical problems, however. Consider the transient analysis of an electromagnetic device which is activated by a voltage (or current) source, such as a transformer, motor, or actuator.

The voltage (current) source for such devices is time-dependent, therefore, one cannot specify *a priori* the numerical value of the current density in the conductive regions of the device, because skin effect and eddy currents cause the current density to vary with time and position within the conductor. The standard finite element procedure, however, requires current density as a known input to the analysis.

In addition, it may be necessary to attach lumped circuit components, such as resistance or inductance, between the voltage (current) source and the region to be modelled by finite elements. The lumped components may represent the internal impedance of the voltage (current) source, or they may be used to approximate the effects of the parts of the device which are outside the region modelled by finite elements. The corresponding transient circuit equations must be coupled with the transient finite element field equations.

Moreover, in the case of the motor or actuator, there are movable mechanical components. Magnetic forces determine the position of these components, and the positions, in turn, affect the magnetic field within the device. Provisions must be made for the transient modelling of such a coupled electro-mechanical system.

Therefore, a method for the proper coupling of transient fields, circuits, and motion must be such that: (a) only terminal voltage (or total terminal current) applied to the device is required as a known input quantity, and total terminal current (terminal voltage) is calculated as an unknown, (b) the transient external circuit equations that model electrical sources and circuit components are coupled to the finite element field equations, and (c) equations for mechanical motion are coupled to the finite element field equations.

This paper presents a method that fulfills these three objectives The method was developed by Istfan [2] and applied by Palma [3]. The following sections detail the method First, the supporting electromagnetic and mechanical theory will be summarized, and the Galerkin formulation of the field equations will be shown Next, the time-discretization of the field, circuit, and mechanical equations will be presented Then the field and mechanical equations will be linearized Finally, the global system of coupled electro-mechanical equations will be assembled and investigated The paper concludes with examples of the application of the proposed method to perform tran

sient analysis of practical coupled electro-mechanical systems.

2 Electromagnetic and Mechanical Theory

The electromagnetic field theory, electric circuit theory, and basic mechanical motion theory, upon which the proposed method is based, are summarized below.

2.1 Summary of Equations

The equations necessary for transient coupling of field, circuit, and motion equations have been derived from basic theory. They are summarized below:

Field Equation

$$\nabla \times \nu \nabla \times A = \sigma \frac{V_b}{\ell} - \sigma \frac{\partial A}{\partial t}$$

Total Current Equation

$$I = \iint_{conductor} \left(\sigma \frac{V_b}{\ell} - \sigma \frac{\partial A}{\partial t} \right) dx\, dy.$$

Series Bar-Coil Equation

$$V_c = \{d_b\}_c^T \{V_b\}_c + L_{ext} \frac{dI_c}{dt} + R_{ext} I_c$$

Parallel Coil Equation

$$V_s = R_s \{1\}^T \{I\}_c + L_s \{1\}^T \left\{ \frac{dI}{dt} \right\}_c + V_c.$$

Mechanical Acceleration Equation

$$m \frac{dv}{dt} + \lambda v = F_{em} - F_{ext}$$

Mechanical Velocity Equation

$$v = \frac{dx}{dt}$$

Note that the first three equations are coupled by the voltage applied to the finite element region, V_b; the second, third, and fourth equations are coupled by the conductor currents, and the first, second, and fifth equations are coupled by magnetic vector potential A.

The portion of the problem to be analyzed with finite elements must be discretized in space, i.e., meshed. The Galerkin method is used to approximate the field and current equations in discretized space. The Galerkin method belongs to a class of approximation techniques called the methods of weighted residuals.

3 Time Discretization

In this section, the field equation, total current equation, circuit equations, and the mechanical equations of motion are discretized in the time domain.

The method of time-discretization used here is based on the following equation:

$$\beta \left\{ \frac{\partial A}{\partial t} \right\}^{t+\Delta t} + (1-\beta) \left\{ \frac{\partial A}{\partial t} \right\}^{t} = \frac{\{A\}^{t+\Delta t} - \{A\}^{t}}{\Delta t}. \quad (1)$$

The value of the constant β determines whether the algorithm is of the forward difference type ($\beta = 0$), backward difference type ($\beta = 1$), or some intermediate type ($0 < \beta < 1$). Note that if $\beta = \frac{1}{2}$, the Crank-Nicholson Method is implemented.

The goal is to solve for $\{A\}^{t+\Delta t}$. The derivatives $\left\{\frac{\partial A}{\partial t}\right\}^{t+\Delta t}$ and $\left\{\frac{\partial A}{\partial t}\right\}^{t}$ are unknown.

4 Linearization

The field equation and the acceleration equations are nonlinear functions of vector potential, A, and/or component displacement, x. These equations must be linearized before they can be combined with the other equations of the system in a general global system matrix equation. The linearization of the field and acceleration equations is accomplished using the Newton-Raphson method.

5 Global System of Equations

The field, circuit, and mechanical equations are now available in a discretized and linearized form. It remains to assemble these matrix equations into a global system of equations describing the entire problem.

5.1 Assembly of the Global System of Equations

From the summary above, it is apparent that, in general, there are five vector unknowns:

$\{\Delta A\}_{k+1}^{t+\Delta t}$. change in vector potential of each node

$\{\Delta V_b\}_{k+1}^{t+\Delta t}$: change in voltage across each bar

$\{\Delta I\}_{c,k+1}^{t+\Delta t}$. change in current in each coil

$\{\Delta V_c\}_{k+1}^{t+\Delta t}$ change in parallel terminal voltage

$\{\Delta x\}_{k+1}^{t+\Delta t}$. change in position of each movable component

The global system matrix equation may then be set up in the form

$$[M]\{f\} = \{N\} \quad (2)$$

or, in expanded form,

$$\begin{bmatrix} M_{1,1} & M_{1,2} & & & M_{1,5} \\ M_{1,2}^T & M_{2,2} & M_{2,3} & & \\ & M_{2,3}^T & M_{3,3} & M_{3,4} & \\ & & M_{3,4}^T & M_{4,4} & \\ M_{1,5}^T & & & & M_{5,5} \end{bmatrix} \begin{Bmatrix} \{\Delta A\} \\ \{\Delta V_b\} \\ \{\Delta I\} \\ \{\Delta V_c\} \\ \{\Delta z\} \end{Bmatrix}_{k+1}^{t+\Delta t} = \begin{Bmatrix} \{N_1\} \\ \{N_2\} \\ \{N_3\} \\ \{N_4\} \\ \{N_5\} \end{Bmatrix} \quad (3)$$

This global system of equations must now be solved for the unknowns, $\{f\}$.

6 Examples

The formulation above has been applied to numerous cases in order to check its validity. Some examples are given below which illustrate the features of coupling fields, circuits, and motion.

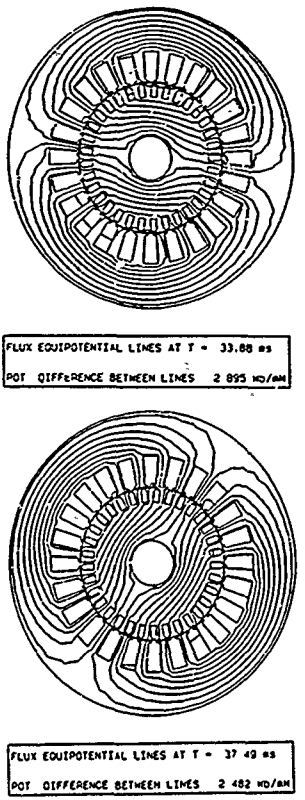


Figure 1: Induction Motor: Equipotential Plots at Different Instants of Time

6.1 Induction Motor

An example is a three phase squirrel cage induction motor. The motor is rated at 5 Hp. In this example the entire motor winding is represented. The inputs are the instantaneous voltages at the three terminals. The current in the windings is unknown. The rotor is free to turn. Each rotor bar is represented as an independent circuit connected to an end ring which has a constant resistance and inductance. The mesh in the air-gap may or may not be remeshed, depending on the distortion of the elements. In any case the remeshing is done such that the number of nodes and elements remains the same. The sequence of plots in Figure 1 shows the motor operating at full load at various positions in the cycle.

7 Summary and Conclusions

A new method for the transient analysis of coupled electromechanical systems has been presented. The principal features of the method are: (a) only terminal voltage (or total terminal current) applied to the device is required as a known input quantity, and total terminal current (terminal voltage) is calculated as an unknown; (b) the transient external circuit equations that model electrical sources and circuit components are coupled to the finite element field equations; and (c) equations for mechanical motion are coupled to the finite element field equations.

The chief attraction of using the method proposed here to develop models for elecro-mechanical devices is that the behavior of the mathematical models is determined by the *same* set of laws that govern the actual devices. The models faithfully reproduce the fact that most actual devices receive input solely from external electrical sources. The interaction of fields, circuits, forces and motion takes place mathematically in the model just as it does physically in the actual device No artificial assumptions about source current density, device inductance, or un-coupled mechanical motion are made.

The proposed method is also attractive in a numerical sense. Care has been taken to arrange system equations in such a way that the global system matrix is sparse and symmetric, thereby allowing the use of efficient storage techniques. In addition, the global system matrix has been made positive definite, so that the system can be solved by a variety of advanced techniques

References

[1] Silvester, P. P. and R. L. Ferrari. *Finite Elements for Electrical Engineers*. Cambridge University Press, Cambridge, 1983.

[2] Istfan, Basim. *Extensions to the Finite Element Method for Nonlinear Magnetic Field Problems*. Ph.D. Thesis, Rensselaer Polytechnic Institute, August 1987.

[3] Palma, Rodolfo. *Transient Analysis of Induction Machines Using Finite Elements*. Ph.D. Thesis, Rensselaer Polytechnic Institute, August 1989.

[4] Coulomb, J. L. "A Methodology for the Determination of Global Electromechanical Quantities From a Finite Element Analysis and its Application to the Evaluation of Magnetic Forces, Torques and Stiffness". *IEEE Transaction on Magnetics*, Vol. MAG-19, No. 6. November 1983. pp. 2514-2519.

[5] Coulomb, J. L. and G. Meunier. "Finite Element Implementation of Virtual Work Principle for Magnetic or Electric Force and Torque Computation". *IEEE Transactions on Magnetics*, Vol. MAG-20, No. 5. September 1984. pp. 1894-1896.

[6] Konrad, A. "The Numerical Solution of Steady-State Skin Effect Problems—An Integrodifferential Approach". *IEEE Transaction on Magnetics*, Vol. MAG-17, No. 1. January 1981. pp. 1148-1152.

[7] Konrad, A. "Integrodifferential Finite Element Formulation of Two-Dimensional Steady-State Skin Effect Problems". *IEEE Transactions on Magnetics*, Vol. MAG-18, No. 1. January 1982. pp. 284-292.

[8] Strangas, Elias G. and Kenneth R. Theis. "Shaded Pole Motor Design Using Coupled Field and Circuit Equations". *IEEE Transactions on Magnetics*, Vol. MAG-21, No. 5. September 1985. pp. 1880-1882.

[9] Strangas, Elias G. "Coupling the Circuit Equations to the Non-Linear Time Dependent Field Solution in Inverter Driven Induction Motors". *IEEE Transactions on Magnetics*, Vol. MAG-21, No. 6. November 1985. pp. 2408-2411.

[10] Hwang, Chang-Chou; S. J. Salon; and R. Palma. "A Finite Element Pre-Processor for Induction Motors Including Circuit and Motion Constraints". *IEEE Transactions on Magnetics*, Vol. MAG-24, No. 6. November 1988. pp. 2573-2575.

POTENTIALS AND GRADIENTS IN A REGION OCCUPIED BY DISTRIBUTED SOURCES IN 2-D FIELDS

Andrzej Dzierzynski, D.Sc., Member, IEEE
The Institute of Electrical Engineering
ul. Pozaryskiego 28
04-703 Warszawa, Poland

Julius Poltz, D.Sc. and Edmund Kuffel, D.Sc. Senior Member, IEEE
The University of Manitoba
Winnipeg, Manitoba
Canada R3T 2N2

Abstract - Potentials and gradients of the electric field in a region occupied by distributed sources can be expressed as some improper integrals. The paper presents formulae for their calculation. The solution can be used in axi-symmetrical cases.

I. INTRODUCTION

Potential and field due to distributed sources can be calculated by integrating the Coulomb's formulae. The integrals become improper for the calculation of potential and field at points within the area occupied by the distributed sources. The region with the sources can be discretized using quadrilateral elements.

In the following it is assumed that the area with the sources is formed by a quadrilateral $P_0P_1P_2P_3$, potential and field being calculated at one of its corners (P_0).

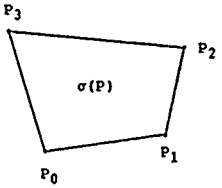


Fig. 1. A quadrilateral element covered with distributed sources of density $\sigma(P)$.

Introducing transformation of a unit square $0 \le u,v \le 1$ onto quadrilateral $P_0P_1P_2P_3$:

$$x = a_x u + b_x v + c_x + d_x uv$$
$$y = a_y u + b_y v + c_y + d_y uv \qquad (1)$$

potential and field at point P_0:

$$V = \frac{-1}{2\pi\varepsilon_0} \int_0^1\int_0^1 \sigma(u,v)\cdot J(u,v)\cdot \ln r(u,v)\ dudv \qquad (2)$$

and

$$E_x = \frac{-1}{2\pi\varepsilon_0} \int_0^1\int_0^1 \sigma(u,v)\cdot J(u,v)\ \frac{x-x_0}{r^2(u,v)}\ dudv \qquad (3)$$

$$E_y = \frac{-1}{2\pi\varepsilon_0} \int_0^1\int_0^1 \sigma(u,v)\cdot J(u,v)\ \frac{y-y_0}{r^2(u,v)}\ dudv \qquad (4)$$

where:

$r^2(x,y) = (x-x_0)^2 + (y-y_0)^2$

$J(u,v)$: Jacobian of the transformation.

The integral corresponding to V and components of the integrals corresponding to E may be represented in a general form as:

$$I = \int_0^1\int_0^1 f(u,v)\cdot h(u,v)\ dudv =$$

$$= \int_0^1\int_0^1 \left[\frac{f(u,v)}{g(u,v)}\ h(u,v) - h(0,0)\right]\cdot g(u,v)\ dudv +$$

$$+ h(0,0) \int_0^1\int_0^1 g(u,v)\ dudv = I_1 + I_2 \qquad (5)$$

where

$$h(u,v) = \frac{1}{2\pi\varepsilon_0}\ \sigma(u,v)\cdot J(u,v)$$

$g(u,v)$: a function that may be integrated analytically and such that the expression

$$p(u,v) = \left[\frac{f(u,v)}{g(u,v)}\ h(u,v) - h(0,0)\right]\cdot g(u,v) \qquad (6)$$

is bounded and integrable over the unit square so that the numerical integration can be applied to calculate the first integral I_1 in (6).

For the calculation of V and E the auxiliary functions $f(u,v)$ and $g(u,v)$ will be discussed separately.

II. CALCULATION OF POTENTIAL V

Let the following functions be defined:

$$f(u,v) = -0.5\cdot\ln\left[(a_x u + b_x v + d_x uv)^2 + (a_y u + b_y v + d_y uv)^2\right] \qquad (7)$$

$$g(u,v) = -0.5\cdot\ln\,(u^2 + v^2) \qquad (8)$$

It can be proved that function p(u,v) obtained by substituting (7,8) in (6) has the following property:

Property P: Function is continuous on the unit square with the exception of point (0,0) in which it is bounded only.

Integral I_2 defined in (6) can be calculated analytically:

$$I_2 = 0.25 \cdot h(0,0) \cdot (6 - \pi - 2 \cdot \ln 2) \qquad (9)$$

III. CALCULATION OF GRADIENT E

Let the following functions be defined:

$$d(u,v) = (a_x u + b_x v + d_x uv)^2 + (a_y u + b_y v + d_y uv)^2$$

$$e(u,v) = (a_x u + b_x v)^2 + (a_y u + b_y v)^2$$

Substituting (1) in (3) yields:

$$E_x = \int_0^1 \int_0^1 h(u,v) \frac{-(a_x u + b_x v + d_x uv)}{d(u,v)} \, du\,dv = a_x I_{ax} + b_x I_{bx} + d_x I_{dx} \qquad (10)$$

where:

$$I_{ax} = \int_0^1 \int_0^1 h(u,v) \frac{-u}{d(u,v)} \, du\,dv \qquad (11)$$

$$I_{bx} = \int_0^1 \int_0^1 h(u,v) \frac{-v}{d(u,v)} \, du\,dv \qquad (12)$$

$$I_{dx} = \int_0^1 \int_0^1 h(u,v) \frac{-uv}{d(u,v)} \, du\,dv \qquad (13)$$

For the calculation of I_{ax} the following functions are used:

$$f(u,v) = \frac{-u}{d(u,v)} \qquad g(u,v) = \frac{-u}{e(u,v)} \qquad (14)$$

Function p(u,v) obtained by substituting (14) in (6) has property P defined in II.

Integral I_2 in (6) can be calculated analytically:

$$I_2 = h(0,0) \cdot S \cdot R \qquad (15)$$

where: $\cdot$ denotes scalar product; vectors R, S are given below:

$$R = \begin{bmatrix} \ln A \\ \ln B \\ \ln (A+B+2C) \\ \operatorname{atan}[(B+C)/D] \\ \operatorname{atan}(C/D) \\ \operatorname{atan}[(A+C)/D] \end{bmatrix} \qquad S^t = \begin{bmatrix} 0 \\ 1/(2A) \\ -1/(2A) \\ -1/D \\ (A-C)/(AD) \\ C/(AD) \end{bmatrix}$$

$$A = a_x^2 + a_y^2 \qquad B = b_x^2 + b_y^2$$

$$C = a_x b_x + a_y b_y \qquad D = a_x b_y - a_y b_x$$

Similar considerations can be applied to calculate I_{bx}. The corresponding integral I_2 in (6) is calculated analytically from:

$$I_2 = h(0,0) \cdot T \cdot R \qquad (16)$$

Vector T is given as:

$$T^t = \begin{bmatrix} 1/(2B) \\ 0 \\ -1/(2B) \\ C/(BD) \\ (B-C)/(BD) \\ -1/D \end{bmatrix}$$

Function under integral I_{dx} (13) also has property P defined in II.

For rectangular elements of constant source density coefficients d_x, d_y and integrals I_1 vanish. Potential and gradient are calculated according to formulae (9) and (10,15,16).

For non-rectangular elements and/or elements of variable source density all the integrals have to be calculated.

IV. NUMERICAL EXAMPLE

Let a quadrilateral element be considered: $P_0(0,0)$, $P_1(1.5, 0)$, $P_2(1, 1)$, $P_3(0.5,1)$

Then the following values of potential and gradient at (0,0) can be calculated:

	V	E_x	E_y
a)	1.988415E-01	(-1.212671,	-7.244254E-01)
b)	2.004412E-01	(-1.212815,	-7.223202E-01)
c)	2.002138E-01	(-1.194719,	-6.990175E-01)

a) - calculated according formulae (5, 9) or (5, 10, 15, 16) and Gauss quadrature of function p(u,v) over the whole element;

b) - as a) but Gauss quadrature with subdivision of the unit square into 64*64 smaller ones;

c) - direct Gauss quadrature over the whole element without taking into account the infinitness of the integrated functions at (0,0)

V. CONCLUSIONS

Differences in calculating potential due to distributed sources within the area occupied by them are rather small. When calculating gradients these differences may reach 3%.

References

1. Lean M.H and Wexler A.:"Accurate Numerical Integration of Singular Boundary Element Kernels over Boundaries with Curvature", Int. Journ. for Num. Meth. in Eng., 1985

FEMAG - AN INTERACTIVE, MENU DRIVEN PACKAGE FOR CAD OF ELECTRICAL MACHINES AND DEVICES

K. Reichert, J. Skoczylas, T. Tärnhuvud

Swiss Federal Institute of Technology (ETH), Departement of Electrical Machines, CH-8092 Zurich, Switzerland

Abstract - FEMAG is the powerful, interactive finite element (FE-) based electromagnetic field analysis system for workstation computers, especially developed for the CAD of electrical machines and devices, with two-dimensional and axisymetrical, static and steady-state nonlinear magnetic field problems. The design of the package is based on industry requirements, where the package is already installed. The paper gives details on the package specifications, on the structure of data and software, and on the program and user interface. Application examples are given at the end.

I. USER REQUIREMENTS

In the last decade a number of 2D and 3D finite element codes have been developed at universities and research centers. Yet the penetration of these methods into a wide range of industries is still not as expected for a number of well known reasons such as lack of user orientation and expertise. Recent developing in computing (workstation, high resolution graphic displays, cost reduction,...) enables the design engineer in industry to solve sophisticated field problems at his desk. However, the designer of new codes for industrial applications, especially for field analysis, should be aware of the following facts:

- Industrial engineers are seldom specialist in FE-method, numerical analysis, operatings systems, computer language.
- The programs are mostly used for basic developements and not in the daily design process.
- The response time should not exeed hours, time required to understand and to use the code should be minimum

The transfer of this situation into specifications for a user- and CAD-oriented FE-package can be summarized as follows:

a) Requirements on problem and field modelling capabilites.

- Problem definition in engineering terms as a local field problem or a device problem considering the coupling of fields and circuits, if required.
- Problem and user oriented postprocessing, i e. direct evaluation of results by means of standard procedures.
- Compatibility of results with those from traditional methods.

b) Requirements on the program interface, structure and handling capabilites:

- Programmed user guidance to minimize learning and relearning times, especially in case of input errors.
- Minimum data input and manual user interaction.
- Communication with the program by means of menues, mouse, tablet, etc.
- Automatic mesh generation, error estimation, adaptive problem oriented mesh refinement.
- CAD interface for geometry input and data base management.
- Multi-task and multi-user enviroment.

These requirements are strongly influencing the theoretical basis, the data structure, the data management, the postprocessing and man-machine interface of the FE-package.

II. DATA STRUCTURE

The minimum and basic data set for FE-calculation consist of

- Node data. Coordinates, potentials, boundary conditions.
- Element data. Type, names of node, material constans, field sources.

As the definition and the input of these basic data is rather labour intensive and offers high failure probabilities, a user oriented data structure is based on:

- Geometrical entities. line, circle, ... segments for the description of the device surfaces and material boundaries.
- Auxiliary entities. Node chains along the geometry, superelements build from closed node chains for the description of the FE-problem.

The node chains are generated manually or automatically based on the geometrical description of the problem The mesh generator determines the basic node and element data set from the auxiliary entities.

User oriented parameter input requires further definitions of entities, especially if material constant or external current or voltage conditions are assigned to a set of superelements. We are using in our package the following additional entities.

- Subregions consisting of a group of superelements having the same material constants, field sources or integral current conditions [3].
- Winding branches consisting of a set of subregions, connected in series and having the same current.
- Winding systems consisting of the winding branches.

This scheme gives the user the possibility, to determine the current density via the winding current, to introduce easy the external current conditions or to evaluate self and mutual inductances, impedances, losses all refered to winding branch. It is especially suited for the analysis of 2D steady-state or transient magnetic field being coupled with the external electric circuits

III. STRUCTURE OF FE-PACKAGE

Program structure: The package FEMAG is a highly interactive, graphic based system for VMS and UNIX environments, being developed in the Dept. of Electrical Machines (ETH)[1]. It consists of four main components:

- interactive preprocessor for geometry definition and semiautomatic mesh generation,
- database with database management system,
- solver with adaptive mesh refinement,
- interactive postprocessing with problem oriented mesh refinements [2].

Interface: Input/output operations use screen-masks and screen-menues. Generally, two levels of mask-menues have been introduced: main menu and submenues (see Fig. 1).

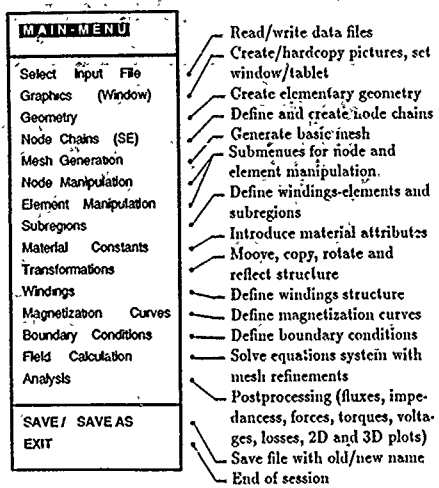


Fig.1. Main menu of FEMAG.

This mode of dialog with the system is uniform, easy to use and almost completely self explanatory. It makes the system easy to learn and to learn it again, which make it especially suited for small industrial design offices, where it may be applied occasionaly. Input data are introduced by means of mouse, keyboard or digital tablet, output uses: graphic display, laser printer (postscript), x-y plotter, graphic printer

Database and database management system: provides easy access to datas by means of a number of well defined subroutines. Structure of FEMAG is modular and data-oriented, therefore easily expandable. Flexibility of data structure has been achieved using special data types (RECORDS) Changes in software involving new group of data, e g. new element attributes, require the change of relevant records only without affecting other parts of the existing software

IV. NUMERICAL EXAMPLES

A number of typical applications of the package for CAD of electrical machines and devices will be discussed during the presentation. One example is shown in Fig. 2 The flux plot in an ac fed squirrel-cage motor at the slip $s = 0.02$ with zero external current conditions for rotor as well, as the calculated and the measured frequency response are shown in Fig. 2.

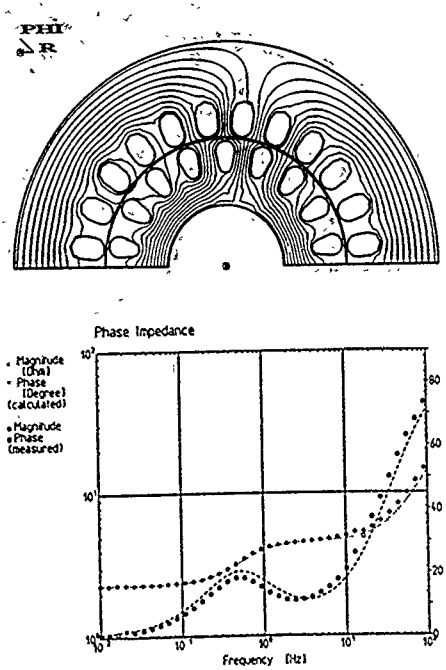


Fig. 2. Flux plot at $t = 0s$ and frequency response of ac fed induction motor.

[1] FEMAG - Reference Manual, Version 5.0, ETH Zürich, 1990.

[2] Tärnhuvud T., Reichert K., Skoczylas J.: Problem-oriented adaptive mesh generation for accurate finite element calculation", IEEE Trans. MAG-90, pp.779-782.

[3] Reichert K., Skoczylas J., Tarnhuvud T. . Eddy current calculations in electrical machines, Numerical solution technique and accuracy problems", Int. Conference on Electrical Machines ICEM-88, pp.59-64.

MECHANICAL STRESSES DISTRIBUTION DUE TO RADIAL SHORT-CIRCUIT FORCES IN TRANSFORMER WINDINGS

ZBIGNIEW WESELUCHA
Institute of Electrical Engineering
04-703 Warsaw, Pozaryskiego 28, POLAND

Abstract - The analytical approach presented in the paper provides a better understanding of the behavior of power transformers under radial short-circuit forces. The inner (LV) winding was taken into account as a multi-layer structure and exemplary results are enclosed.

INTRODUCTION

The current carrying conductors of the transformer windings are situated in the region of magnetic leakage flux and experience mechanical forces. The excessive increase in KVA ratings of large power transformers in the last years accompanied by a parallel increase in the forces generated in the windings under short-circuit conditions, are the main reasons of mechanical damages of transformers. Statistic documents report an increasing number of power transformers failures under mechanical forces produced by fault currents [5,8].

In concentric winding the axial component of the leakage flux produces radial forces (Fig.1).

The radial forces may cause the well-known buckling phenomenon of inner (LV) windings. Previous concepts of calculating those mechanical damages were based on the assumption of the average value of the stress in each conductor [3] and were far from expected ones. The proposition placed in the paper increases the possibility of the mathematical model of transformer coil and lets to observe the extreme values of stresses, strains and displacements in each conductor. The circumferential component of stress in conductor expresses mechanical withstanding of a coil due to radial forces

LIST OF SYMBOLS

- B_z - axial component of the magnetic flux density
- E - modulus of elasticity
- F_r - radial component of the density of electromagnetic force
- J - current density in conductor
- r,θ - polar coordinates
- R_z, R_w - outer and inner radii of a winding
- u,v - radial and circumferential components of components of displacement
- $\delta = \begin{cases} 1 & \text{- for conductor material} \\ 0 & \text{- for insulation material} \end{cases}$
- $\varepsilon_r, \varepsilon_\theta$ - radial and circumferential components of strain
- γ - tangential component of strain
- κ - stiffeness of support
- ν - Poisson's ratio
- σ_r, σ_θ - radial and circumferential components of stress
- τ - tangential component of stress

FORMULATION

Since the generated forces are electromagnetic in origin they are termed here as the "electromagnetic" forces. The radially-acting components of the forces in a transformer windings are easily and accurately calculated by elementary methods [2,7,8]. From the theoretical point of view we say here about "weak-coupled" electromagnetic and mechanical fields.

The electromagnetic force density is determined as follows [7]

$$F_r = J * B_z \qquad (N/m^3) \qquad (1)$$

Transformer coil is a multi-layer ring consisting of conductor and insulation materials (Fig.1 B)

The considerations for the inner winding may be reduced to the one coil chosen from the middle of the winding

We call a "zone" each conductor and insulation layer of the multi-layer coil structure.

The inner winding (LV) experiences a force acting inward tending to crush or collapse it [8]. It is especially difficult problem since the windings are supported from the core by the distance spacers round the circumference. Failure may occur by bending in of the conductors between supports producing a characteristic star-shape.

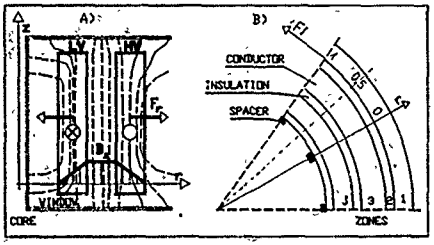


Fig.1. A) Axial component of magnetic flux density and radial forces in simple concentric windings,
B) The multi-layer transformer coil structure.

All considerations are based on the assumption that the stress does not exceed the proportional limit; this mean that strain must be proportional to stress, in accordance with Hooke's law (2). The material of the ring must be elastic, thus capable of sustaining stress without permanent deformation.

BASIC EQUATIONS

2-D Hooke's Law (in polar coordinates)

$$\varepsilon_r = \frac{1}{E}[\sigma_r - \nu\sigma_\theta], \quad \varepsilon_\theta = \frac{1}{E}[\sigma_\theta - \nu\sigma_r], \quad \gamma = \frac{E}{2(1+\nu)}\tau \qquad (2)$$

Cauchy's relationships (strains-displacements):

$$\varepsilon_r = \frac{\partial u}{\partial r}, \quad \varepsilon_\theta = \frac{1}{r}\frac{\partial w}{\partial \theta} + \frac{u}{r}, \quad \gamma = \frac{1}{r}\frac{\partial u}{\partial \theta} + \frac{\partial w}{\partial r} - \frac{w}{r} \qquad (3)$$

From the theory of elasticity we get 2-D Kirchhof's problem which is described by two differential equations of static equilibrium as follows [1]

$$\left.\begin{aligned} &\frac{\partial \sigma_r}{\partial r} + \frac{1}{r}\frac{\partial \tau}{\partial \theta} + \frac{\sigma_r - \sigma_\theta}{r} + \delta F_r = 0 \\ &\frac{\partial \tau}{\partial r} + \frac{2\tau}{r} + \frac{1}{r}\frac{\partial \sigma_\theta}{\partial \theta} = 0 \end{aligned}\right\} \qquad (4)$$

BOUNDARY CONDITIONS:

1. continuity of stresses and displacements for each "j" and "j+1" zones

$$\left.\begin{aligned} \sigma_r^j &= \sigma_r^{j+1} \\ \tau^j &= \tau^{j+1} \\ u^j &= u^{j+1} \\ v^j &= v^{j+1} \end{aligned}\right\} \quad (5a)$$

2. zero stress condition for $r = R_z$

$$\left.\begin{aligned} \sigma_r &= 0 \\ \tau &= 0 \end{aligned}\right\} \quad (5b)$$

3. zero stress condition for $r = R_w$

$$\tau = 0 \quad (5c)$$

4. dual stress-displacement condition for $r = R_w$ [7] representing a contact problem

$$\left.\begin{aligned} u &= \frac{1}{\kappa}\hat{\sigma}_r \quad \text{- for support region} \\ \sigma_r &= 0 \quad \text{- outside the support} \end{aligned}\right\} \quad (5d)$$

Substituting (2) and (3) into (4) we obtain differential equation for radial displacements u, the solution of which and boundary conditions (5) allow us to reduce the contact problem to Fredholm integral equation of the first kind. The solution of the integral equation is the base of the numerical algorithm.

RESULTS

The following parameters of the coil were taken for the calculations:

5 conductors in a coil, inner radius R_w = 650 mm,
36 spacers around circumference,
conductor width 1.6 mm, insulation width 0.3 mm,
spacer width 20 mm,
Poisson's ratio for conductor $\nu = 0.32$,
modules of elasticity E: - conductor $1.125 \cdot 10^5$ MPa,
- insulator 250 MPa,
- spacers 350 MPa.

The maximum electromagnetic force has been assumed as the unit quantity.

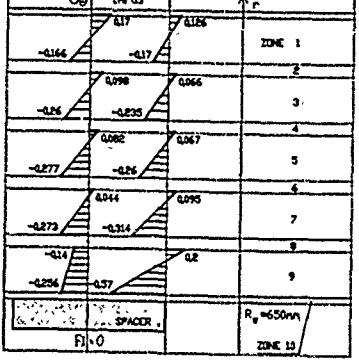


Fig.2 Distribution of circumferential stresses in the support region (the middle and the end of spacer)

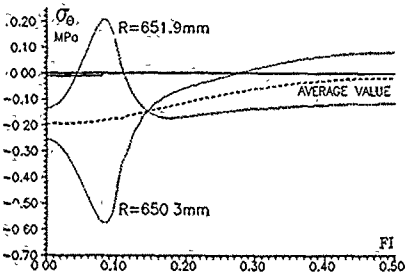


Fig.3. Distribution of an average and extreme values of stresses σ_θ in the inner conductor of a coil.

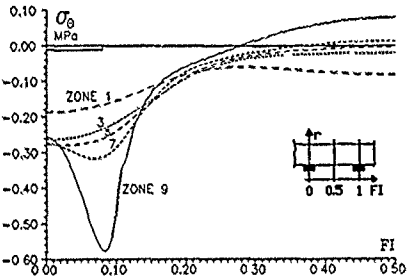


Fig.4. Distribution of the extreme values of circumferential stresses σ_θ in every conductor of a coil.

CONCLUSION

It has been shown (Fig.2,3) that the mechanical "overstresses" occurring in conductors of a coil near the end of the spacers are even few times greater then the average value of stress. The extremely hard conditions appear in the inner conductor of a coil supported on the stiff spacers.

REFERENCES

[1] Huber H.T.: Teoria sprężystości, vol.I, PWN, 1954
[2] Jezierski E.: Transformatory, WNT, 1975
[3] Hilman L., Lurie S. The Short-Circuit Strength of the Inner Transformer Windings, Elektritchestvo, No 3, 1968 (in Russian).
[4] Noras E.T.: Mechanical Strength of Power Transformers in Service, Proc.IEE, Vol 104A, 1957.
[5] Patel M.R.. Dynamic Response of Power Transformers under Axial Short-Circuit Forces, IEEE Trans on PAS vol PAS-94, no.2, 1975.
[6] Timoshenko S , Woinowsky-Krieger S., Theory of Plates and Shells, McGraw-Hill, 1959.
[7] Turowski J.. Elektrodynamika techniczna, WNT, 1968
[8] Waters M . The Short-Circuit Strength of Power Transformers, Mcdonald's, London, 1966.
[9] Wesełucha Z.. Stress Field Distribution Analysis of the Inner Transformer Winding Due To Radial Forces, Dr Diss (in Polish), 1990

A GENERAL SCHEME FOR CUTTING THE MAGNETIC SCALAR REGION IN MULTIPLY CONNECTED 3D EDDY CURRENT PROBLEMS

P. J. Leonard and D. Rodger
School Electrical Engineering
University of Bath, UK

Abstract — *The efficient solution of eddy current problems can be achieved using the Aψ formulation. This does however require cuts to be made in the magnetic scalar region whenever the conductor is multiply connected. This paper presents a method for modelling cuts for a variety of topologies including interlinked conducting circuits.*

I. INTRODUCTION

In a previous paper [1] a scheme was described for cutting the magnetic scalar potential over the spanning surface of a hole in a conductor. The scheme was limited to surfaces which did not intersect themselves. In this paper we generalises the scheme to allow intersecting surfaces and folds.

II. THEORY

A. *Modelling a single hole*

The Aψ formulation is used to model eddy current problems. The conductor is modelled using the magnetic vector potential, whilst surrounding non-conductors modelled using the magnetic scalar potential. The usual finite element approximation for ψ is a continuous function. Reference [1] describes how the representation can be modified to allow a jump in the value of ψ as we pass through a surface spanning a hole.

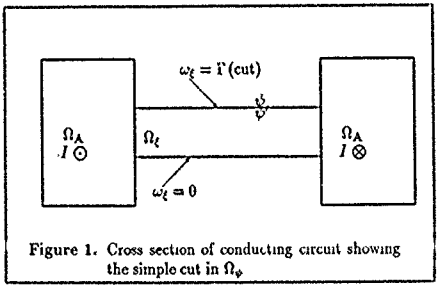


Figure 1. Cross section of conducting circuit showing the simple cut in Ω_ψ

The definition of the magnetic scalar is modified in a layer of elements that spans the hole (see figure 1):

$$H = -\nabla\omega_\xi\psi_\xi - \sum \nabla\omega_i\psi_i \qquad (1)$$

The ω_i are the normal basis functions for a continuous scalar potential; ω_ξ is an additional discontinuous basis function, that lives within the layer of elements spanning the hole.

The additional basis function is defined to be zero on one side of the layer and unity on the other. With this new basis function the value of ψ jumps as it passes through the $\omega_\xi = 1$ surface,

$$\psi' = \psi + \psi_\xi \qquad (2)$$

The finite element model follows the usual development. However we have an extra equation corresponding to the new unknown,

$$-\int_{\Gamma_{\xi A}} (\nabla\omega_\xi \times A)\cdot\hat{n}\, d\Gamma + \int_{\Omega_\xi} \mu\nabla\omega_\xi\cdot\nabla\psi\, d\Omega = 0 \qquad (3)$$

In this equation ω_ξ can be regarded as a test function which (weakly) ensures that the line integral of A around the cut is equal to the magnetic flux flowing through the surface.

It is easily seen that any number holes can be modelled using this technique providing the surfaces do not interfere with each other.

B. *Conducting bar piercing cut*

Figure 2 shows a slightly more complex situation, there is a conducting bar which interferes with the cutting surface.

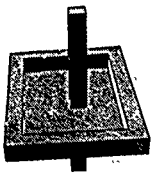


Figure 2. Showing a loop with a bar which pierces the cutting surface

In fact this case does not need any special treatment. The reduction of the integral in the Ω_ξ region should be equal to the increase of the integral over the $\Gamma_{\xi A}$ interface (equation 3.

C. *Interconnected circuits*

If two circuits link each other as shown in figure 3, then the two basis functions for the cuts interfere. However the principles remain the same, we just have two extra equations. Each equation will see a contribution from the other through the definition of ψ

$$\psi = \omega_{\xi 1}\psi_{\xi 1} + \omega_{\xi 2}\psi_{\xi 2} + \sum \omega_i\psi_i \qquad (4)$$

Then the equation for the test function, $\omega_{\xi 1}$, is:

$$-\int_{\Gamma_{\xi 1 A}} (\nabla\omega_{\xi 1} \times \mathbf{A}) \cdot \hat{\mathbf{n}}\, d\Gamma + \int_{\Omega_{\xi 1}} \mu \nabla\omega_{\xi 1} \cdot \nabla\psi d\Omega = 0 \quad (5)$$

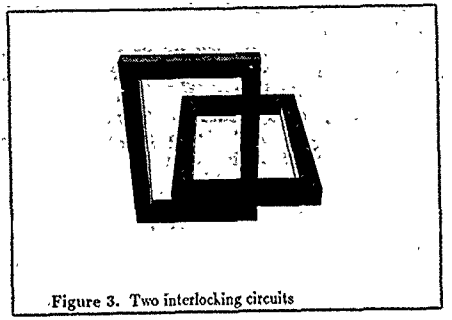


Figure 3. Two interlocking circuits

D. *Self intersecting or folded cutting surfaces*

Some shapes of circuit have holes which are more difficult to cut. For example a helical winding is topologically equivalent to a doughnut but a cutting surface which does not interfere with itself is more difficult to construct.

Allowing the surface to fold and intersect itself helps the construction of the cutting surface but requires a modification of the basis function ω_ξ. The correct strategy is to add the contributions when these folds or intersections occur. In practice it is easier to construct the cutting surface which makes the problem simply connected and then deduce the require jumps [2].

For example consider the loop of conductor shown in figure 4, and the cutting scheme shown in figure 5.

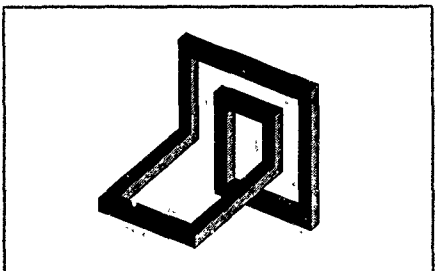


Figure 4. Loop of conductor

The outermost surface requires a jump equal to the current in the loop whilst the inner surface requires twice this amount The basis function required to produce this jump is illustrated in figure 6. The integers are the the nodal values of the basis function ω_ξ within a particular region, thus the value of the basis at a given node will depend on which element we are looking at. Note that this could be regarded as the sum of two overlapping surfaces.

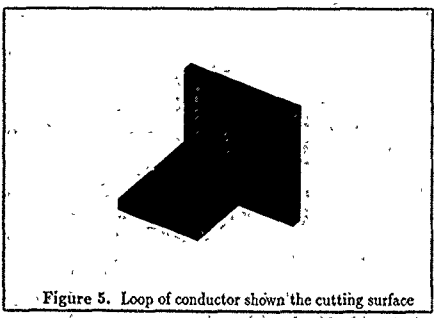


Figure 5. Loop of conductor shown the cutting surface

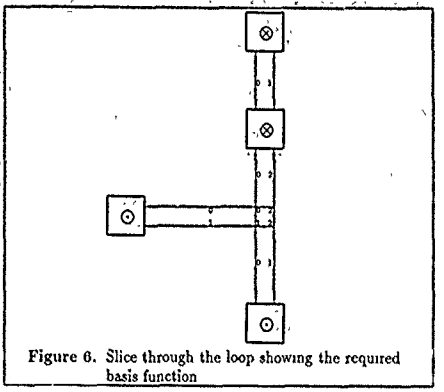


Figure 6. Slice through the loop showing the required basis function

III. CONCLUSION

We have presented a scheme for cutting the magnetic scalar region in multiply connected eddy current problems and given a simple example of it's application. The scheme can be used even when cutting surfaces overlap.

References

[1] P.J.Leonard and D.Rodger. A new method for cutting the magnetic scalar potential in multiply connected eddy current problems *IEEE Trans Mag.*, 25(5), Sept 1989.

[2] M. L Brown. Scalar potentials in multiply connected regions. *IJNME*, 20 665-680, 1984.

Three-dimensional Electromagnetic Computations in Quasi Axisymmetric Structures

R. Albanese† G. Rubinacci‡

Introduction. Several numerical procedures and related computer codes are now available for the analysis of electromagnetic (e.g. magnetostatic and eddy current) problems in three-dimensional geometries. However, there are a number of applications for which fully three-dimensional models can hardly be used. An example of this kind is furnished by the analysis of a plasma discharge in a Tokamak. In such devices the hypothesis of axisymmetry is acceptable and even mandatory for the plasma. On the other hand, the hypothesis is not directly applicable to the patterns of the eddy currents and electromagnetic fields in the metallic structures, due to their segmentation along the toroidal direction. Nevertheless, the interest is often focused on the plasma region. Here, the electromagnetic fields produced by eddy and magnetising currents localized in the metallic structures in the presence of typical axisymmetric excitations (currents flowing in poloidal field coils, plasma motion and disruptions) are substantially axisymmetric. In this case, it is possible to approximate the quasi-axisymmetric metallic structures by means of equivalent axisymmetric models connected to a fictitious electric network. Parameters (resistances and inductances) of this network and material properties of the axisymmetric model can be selected such as to mantain the input/output behavior as close as possible to that of the three-dimensional system.

Three-dimensional model. The eddy current problem in a three-dimensional conducting region $\mathcal{V}$ can be described by expanding the eddy current density as $\mathbf{J}(\mathbf{x},t) = \sum_{k=1}^{n} I_k(t)\mathbf{J}_k(\mathbf{x})$, where the basis functions $\mathbf{J}_k(\mathbf{x})$ are constrained to be solenoidal in $\mathcal{V}$ with $\mathbf{J}_k(\mathbf{x}) \cdot \mathbf{n} = 0$ on $\partial\mathcal{V}$. For nonmagnetic conductors, the time evolution of the coefficients $I_k(t)$ can be obtained by means of the following linear initial value problem [1]:

$$\underline{L}\,\dot{\underline{I}} + \underline{R}\,\underline{I} = \underline{V} \tag{1}$$

with $L_{ik} = \frac{\mu_0}{4\pi}\int_{\mathcal{V}}\int_{\mathcal{V}} \frac{\mathbf{J}_i(\mathbf{x})\,\mathbf{J}_k(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d\mathcal{V}\,d\mathcal{V}'$, $R_{ik} = \int_{\mathcal{V}} \mathbf{J}_i(\mathbf{x})\,\eta\mathbf{J}_k(\mathbf{x})\,d\mathcal{V}$ and $V_i = \int_{\mathcal{V}} \mathbf{J}_i(\mathbf{x}) \cdot \mathbf{E}_e(\mathbf{x},t)\,d\mathcal{V}$. Here μ_0 is the vacuum permeability, η the resistivity and $\mathbf{E}_e$ the externally applied electric field. The strict analogy with a lumped parameter approach is evident: $\underline{L}$ and $\underline{R}$ play the roles of the inductance and resistance matrices, respectively, whereas $\underline{I}$ and $\underline{V}$ play the roles of currents and applied voltages.

In basically axisymmetric systems like the tokamak reactors, it is interesting to analyze the influence of the quasi-axisymmetric structures on the mutual interaction between axisymmetric conductors. In this case, the applied electric field is usually given by a set of axisymmetric currents I_{ek} (input quantities):

$$\underline{V} = -\underline{M}_e\,\dot{\underline{I}}_e \tag{2}$$

where M_{eik} is the magnetic flux linked with I_{ek}, due to the unit vector current density shape function $\mathbf{J}_i$; the relevant output quantities to be computed are generally the magnetic fluxes $\underline{\psi}_0$ linked with a number of axisymmetric coils:

$$\underline{\psi}_0 = \underline{M}_0 \cdot \underline{I} + \underline{M}_{0e} \cdot \underline{I}_e \tag{3}$$

M_{0eik} is the flux produced by the unit axisymmetric current I_{ek} and linked with the circumference Γ_i, whose center is on the toroidal axis of symmetry; M_{0ik} has a similar meaning but is due to the degree of freedom I_k associated with the shape function $\mathbf{J}_k$. Therefore, in complex notation, the system can be put in the form:

$$(s\underline{L} + \underline{R}) \cdot \hat{\underline{I}} = -s\underline{M}_e \cdot \hat{\underline{I}}_e \tag{4}$$

$$\hat{\underline{\psi}}_0 = \underline{M}_0 \cdot \hat{\underline{I}} + \underline{M}_{0e} \cdot \hat{\underline{I}}_e \tag{5}$$

where $\hat{}$ indicates a variable in the s domain.

Quasi-axisymmetric model. Toroidally continuous structures can well be schematized as axisymmetric conductors having equivalent sections and resistivities. This is for instance the case of the vacuum vessel of a tokamak. However, there are structures like the tokamak first walls which consist of a number of toroidally insulated sectors (see Figure 1). As suggested in [2,3], structures of this kind can be schematized as a set of axisymmetric coils connected to an external network. The parameters (inductances and resistances) of the branches of this fictitious network can be selected to approximate the effects of the nontoroidal currents. The behavior of such a system can be described by the following equations:

$$(\underline{Y}^{-1} + s\tilde{\underline{L}} + \tilde{\underline{R}}) \cdot \hat{\underline{I}}_{ax} = -s\tilde{\underline{M}}_e \cdot \hat{\underline{I}}_e \tag{6}$$

$$\hat{\underline{\psi}}_0 = \tilde{\underline{M}}_0 \cdot \hat{\underline{I}}_{ax} + \underline{M}_{0e} \cdot \hat{\underline{I}}_e \tag{7}$$

Here $\tilde{}$ indicates the matrices due to the axisymmetric contributions and $\hat{I}_{ax}$ indicates the axisymmetric independent

*Research partially supported by Euratom and by Ministero dell'Università e della Ricerca Scientifica e Tecnologica

†Istituto di Ingegneria Elettronica, Università di Salerno, Italy
‡Dipartimento di Ingegneria Industriale, Università di Cassino, Italy

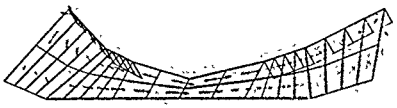


Figure 1: Current distribution in one half of a blanket/first wall module.

currents. $\underline{Y}$ is the admittance matrix of the external fictitious network. The topology of this network is chosen such as to have a close relationship with a phisical interpretation of the circuit parameters involved. These unknown parameters can be initially computed using approximate analytical expressions, which in case of a tokamak first wall are reported in [2,4]. Of course the I/O behaviors of the 3D system:

$$\underline{H}_{3D}(s) = -\underline{M}_0 \cdot (s\underline{L} + \underline{R})^{-1} \cdot s\underline{M}_e \qquad (8)$$

and of the approximate 2D quasi axisymmetric system:

$$\underline{H}_{2D}(s) = -\tilde{\underline{M}}_0 \cdot (\underline{Y}^{-1} + s\tilde{\underline{L}} + \tilde{\underline{R}})^{-1} \cdot s\tilde{\underline{M}}_e \qquad (9)$$

will result to be different. However, their difference can be reduced by using a procedure similar to the Levy's Method [5]. In fact, the unknown coefficients of the polinomials at the numerator and denominator of H_{2D} ($H_{2D}(s) = B(s)/A(s)$) can be computed as the minimum of the quadratic form $U = \sum_{i=1}^{N} E_i^* E_i$ where $E_i = A(j\omega_i)H_{3D}(j\omega_i) - B(j\omega_i)$, ω_i are a number of suitable frequencies and $\cdot^*$ denotes adjoint. Of course, the relationship among coefficients and circuit parameters is non linear and can be obtained by a classic procedure of circuit synthesis. In a more direct numerical way, resistances R_{fj} and inductances L_{fj} of the fictitious external network can be tuned in order to minimize the functional:

$$\sum_{i=1}^{n} \frac{|\underline{H}_{3D}(j\omega_i) - \underline{H}_{2D}(j\omega_i)|^2}{|\underline{H}_{3D}(j\omega_i)|^2} \qquad (10)$$

where ω_i are again a number of suitable frequencies. R_{fj} and L_{fj} should be positive, so that, in order to have an unconstrained minimization problem, they shoud be rewritten as $R_{fj} = R_{0fj}\alpha_j^2$, $L_{fj} = L_{0fj}\beta_j^2$. The problem is nonlinear. However, there is the advantage that the physical interpretation is not lost, and that, apart the I/O behavior, other properties are somehow preserved by the quasi-axisymmetric model.

Results and conclusions. As an example, we studied the blanket/first wall geometry described in Figure 1. Due to the symmetries of the system, the 3D numerical model (72 elements, 62 degrees of freedom) is limited to one half of one of the 48 modules. The conductors carrying the currents $\underline{I}_e$ are symmetrically located with respect to the equatorial plane. Two possible inputs are considered: the first, symmetric, produces on the equatorial plane a vertical field B_z, the second, antisymmetric,

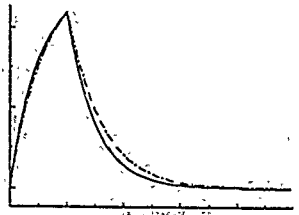


Figure 2: B_R as a function of time. The curves refer to the output of 3D model (——), 2D approximate model (-·-·-) and 2D optimised model (- - -).

gives at the same position a radial field B_R. Accordingly, the outputs are the flux linked with the plasma and two flux combinations giving the field $B_z \approx \Delta\psi/2\pi R\Delta R$ and $B_R \approx -\Delta\psi/2\pi R\Delta z$.

The 2D system is made out of 16 axisymmetric conductors, properly connected to an external fictitious network, as previously described.

The antisymmetric output (the field B_R due to an antisymmetric input linearly rising to 1 MA in 10 ms and then constant), is shown in Figure 2, where a comparison is made among the 3D case, the approxmate 2D case whose network parameters have been analitycally estimated, and the optimized 2D case.

In this simple case, it can be seen that the axisymmetric model works in a satisfactory way also without the optimization procedure. Of course, more complex real cases cannot be handled with a similar accuracy without adopting a systematic approach like the one here described. The I/O behavior is preserved with a little effort (unconstrained minimization of a nonlinear functional). However, the physical interpretation is not lost. Other local and global quantities (other than those used in the optimization process) can be obtained with sufficient accuracy. In this respect, the optimization plays the role of a tuning of the parameters of the fictitious network. Finally, it is worthwhile to notice that the relevant 3D transfer function can also be obtained by means of a set of measurements.

References

[1] R. Albanese and G. Rubinacci, *Integral formulation for 3D eddy current computation using edge elements.* IEE Proc., Vol.135(7), Pt.A, 1988, pp. 457-462.

[2] R. Albanese, *Analysis of the plasma equilibrium evolution in the presence of circuits and massive conducting structures.* 15th Symposium on Fusion Technology, Utrecht, Sept. 1988, Elsevier 1989, pp. 281-286.

[3] R. Albanese, E. Coccorese and G. Rubinacci, *Plasma modelling for the control of vertical instabilities.* Nuclear Fusion, Vol 29, No 6, 1989, pp. 1013-1023.

[4] E. Coccorese and F Garofalo, in *Tokamak Start-Up* Knoepfel ed., E. Majorana series, Vol 26, Plenum Press, New York, 1986, p.337

[5] G. C Goodwin and R. L. Payne, *Dynamic System Identification. Experiment Design and Data Analysis* Academic Press, 1977.

3D COMPUTATIONS IN ELECTROMAGNETICS

O. BIRO AND K. PREIS
Graz University of Technology, Kopernikusgasse 24, A-8010 Graz, Austria

Abstract - Three-dimensional computations of electromagnetic fields are presented with the aid of the method of finite elements. Magnetostatic, eddy current and microwave fields are treated. Various formulations are discussed with some advantages and disadvantages pointed out.

INTRODUCTION

The differential equations of electromagnetics can be set up in terms of various variables; it is possible to use the field quantities directly, or to introduce potential functions. In the course of their numerical solution, the differential equations are replaced by sets of algebraic equations. If nodal finite elements [1] are used to set up these equations, the potentials are advantageous, since they are continuous. Recently, vectorial or edge elements are also frequently employed [2,3], they fit well to a representation by field vectors whose tangential components only are continuous.

For three-dimensional computations, the numerical stability of the method is extremely important. Due to the large number of degrees of freedom, iterative methods are bound to be used for the solution of the equations, the most frequent one being the method of conjugate gradients [4]. Unless the uniqueness of the variables is ensured, the number of iterations needed becomes unbearably high.

In the following discussion of various 3D computations, the authors' experience with respect to the above two points will be outlined.

MAGNETOSTATIC FIELDS

Magnetostatic fields can be represented with the aid of a reduced scalar potential [5], i.e. as the sum of an impressed field satisfying Ampere's law and of the gradient of a scalar potential. Nodal finite elements suit well for the approximation of this scalar. Serious cancellation errors occur in highly permeable parts where the impressed field calculated accurately by Biot-Savart's law is several orders of magnitudes higher than the resultant field obtained by subtracting the gradient of the reduced scalar potential, an operation involving numerical differentiation. The most widely used method to overcome this difficulty is to set the impressed field to zero in regions with high permeability, thus arriving at a total scalar potential here [5]. Another possibility is to compute the impressed field under the assumption of infinite permeability [6]. In both cases, the discontinuities of the impressed field give rise to additional interface conditions to be satisfied by the scalar potential. The authors have recently experimented successfully with a continuous impressed field represented with the aid of edge elements [7]. A similar method was suggested in [8].

Recent results of a TEAM Workshop problem by several groups around the world by widely differing methods [9,10] have pointed out that scalar potential methods may yield too high field values in ferromagnetic regions. The problem consists of iron channels around a racetrack shaped coil. The flux density is shown in Fig. 1 in one eighth of the model as obtained by total and reduced scalar potentials by C. Magele [10]. The results seem completely acceptable, but the value of the flux density in the iron plates is about 10 percent higher than the measured value even if the discretization is extremely fine.

An alternative to the scalar potential representation is to use a magnetic vector potential. In case of nodal elements, ill-conditioning occurs if no gauging is used [11]. The Coulomb gauge can be enforced on the vector potential to ensure its uniqueness and thus numerical stability [12], however, too low flux densities are obtained in ferromagnetic parts [11]. This problem has been overcome by the authors by freeing the normal component of the vector potential over iron/air interfaces [11,7]. The alternative of employing edge elements to represent the vector potential have yielded encouraging results (see results by T. Nakata, [9]), but it seems that the numerical stability becomes inferior and it can be little improved by gauging [13].

It seems that the computation of magnetostatic fields in three dimensions is by no means a problem that can be considered to have been solved long ago.

EDDY CURRENT FIELDS

Due to the coupling between the electric and magnetic field, eddy current fields cannot be described by a scalar. Various continuous potentials can be introduced to represent them, in the most general case both a vector and a scalar potential are necessary in eddy current carrying conductors. One possible pair is a magnetic vector potential and an electric scalar potential (A,V), the other is a current vector potential and a magnetic vector potential (T,ψ). The static magnetic field in the surrounding nonconducting region can again be described either by a magnetic scalar potential or by a magnetic scalar potential. All these formulations in conductors and nonconductors can be coupled to each other in a symmetric way [14]. In order to achieve numerical stability, it is necessary to gauge the vector potentials. A possibility is to enforce the Coulomb gauge on them [12].

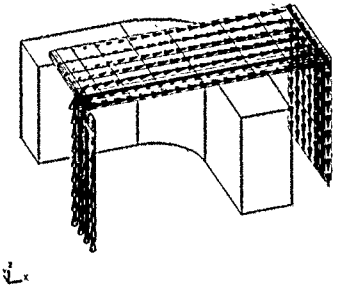


Fig. 1 Flux density distribution in ferromagnetic plates

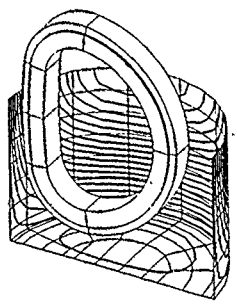


Fig. 2 Eddy current density in a liquid nitrogen shield

In some problems, the eddy current carrying conductors are thin sheets with negligible variation of the eddy current density in the direction normal to their surface and with the normal component of the current density practically zero. In this case the use of volume elements to describe the eddy currents leads to numerical problems due to the large differences in element sizes in the vicinity of the conductors. A possibility is to model the conductors by sheets of zero thickness and to describe the eddy currents by a single component current vector potential. A method to couple this stream function to a scalar potential in the surrounding nonconducting region has been presented in [15]. The authors have recently developed a method to connect the above description with a vector potential formulation of the surrounding static field [16]. This method has been applied to the computation of the transient eddy current field of a large superconducting coil in an experimental setup for future TOKAMAK reactors [17]. The distribution of the eddy current density in a liquid nitrogen shield around such a D-shaped coil due to a discharge is shown in Fig. 2

3D CAVITIES

In the analysis of 3D cavities, the problem is to find the frequencies and the corresponding field patterns that can exist without external excitation. In contrast to the previous deterministic problems, this leads to an eigenvalue problem. The finite element computation of such problems has long been plagued by so called spurious modes which are nonphysical solutions that approximate no real mode. They are now recognized as approximations of gradient fields corresponding to the zero frequency eigenvalue that has infinite multiplicity unless the vector variable used is gauged properly [18].

One possibility to eliminate spurious modes is to employ nodal elements but enforce the Coulomb gauge on the vector potential with a scalar potential also used [19]. This method leads to as many zero eigenvalues as there are unknown scalar potential values. The second mode of a rectangular cavity partially filled by a dielectric slab [20] is illustrated in Fig. 3 by the distribution of the electric field.

Spurious modes can also be avoided by using edge elements and a field quantity as system variable. In this way, the space of gradient fields is projected to a subspace of finite dimensions in the space spanned by the edge elements. Therefore, similarly to the above method used in conjunction with nodal elements, the number of zero eigenvalues is known in advance, with the rest of the modes approximating physical solutions [18,21].

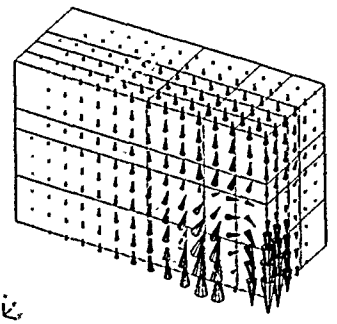


Fig. 3. Electric field distribution in a dielectric loaded cavity

CONCLUSION

Some formulations for the finite element computation of electromagnetic fields in three dimensions have been reviewed. The importance of the appropriate choice of the system variables, of the gauge conditions and of the elements used has been pointed out.

REFERENCES

[1] O.C. Zienkiewicz, *The Finite Element Method*, 3rd ed. New York, NY: McGraw-Hill, 1977.

[2] R. Albanese, G. Rubinacci, "Solution of three dimensional eddy current problems by integral and differential methods", *IEEE Trans. on MAG.*, vol. 24, pp. 98-101, 1988.

[3] A. Kameari, "Calculation of transient 3D eddy current using edge-elements", *IEEE Trans. on MAG.*, vol. 26, pp. 466-469, 1990.

[4] D.S. Kershaw, "The incomplete Cholesky-conjugate gradient method for iterative solution of systems of linear equations", *J. Comput. Phys.*, vol. 26, pp. 43-65, 1978.

[5] J. Simkin, C.W. Trowbridge, "On the use of the total scalar potential in the numerical solution of field problems in electromagnetics", *Int. J. Numer. Meth. Eng.*, vol. 14, pp. 423-440, 1979.

[6] I.D. Mayergoyz, M.V.K. Chari, J. D'Angelo, "A new scalar potential formulation for three-dimensional magnetostatic problems", *IEEE Trans. on MAG.*, vol. 23, pp. 3889-3894, 1987.

[7] K. Preis, I. Bardi, O. Biro, C. Magele, G. Vrisk, K.R. Richter, "Different finite element formulations of 3D magnetostatic fields", submitted to *COMPUMAG'91, Sorrento, 7-11 July, 1991*

[8] J P. Webb, B. Forghani, "A single scalar potential method for 3D magnetostatics using edge elements", *IEEE Trans. on MAG.*, vol. 25, pp. 4126-4128, 1989.

[9] Y. Crutzen, N.J. Diserens, C.R.I. Emson, D. Rodger (Ed.), "European TEAM Workshop and International Seminar on Electromagnetic Field Analysis", Oxford, England, 23-25 April 1990.

[10] K.R. Richter, W.M. Rucker, O. Biro (Ed.), "4th International IGTE Symposium and European TEAM Workshop", Graz, Austria, 10-12 October, 1990.

[11] K. Preis, I. Bardi, O. Biro, C. Magele, W. Renhart, K.R. Richter, G. Vrisk, "Numerical analysis of 3D magnetostatic fields," 4th Biennial IEEE Conf. Electromagn. Field Comp., Oct. 22-24, 1990 Toronto, Ontario, Canada

[12] O. Biro, K. Preis, "On the use of the magnetic vector potential in the finite element analysis of 3-D eddy currents", *IEEE Trans. on MAG*, vol. 25, pp. 3145-3159, 1989.

[13] A. Kameari, "Study on 3-D eddy current analysis using FEM", 4th Int. IGTE Symp. and Europen TEAM Workshop, Graz, Austria, 10-12 Oct. 1990.

[14] O. Biro, K. Preis, "Finite element analysis of 3-D eddy currents", *IEEE Trans on MAG*, vol. 26, pp. 418-423, 1990.

[15] D. Rodger, N. Atkinson, "A finite element method for 3D eddy current flow in thin conducting sheets", *Proc. IEE, Pt. A*, vol. 135, pp. 369-374, 1988.

[16] O. Biro, R. Heller, P. Komarek, W. Maurer, K. Preis, K.R. Richter, "FEM calculations of eddy current losses and forces in thin conducting sheets of test facilities for fusion reactor components", submitted to *COMPUMAG'91, Sorrento, 7-11 July, 1991*

[17] O. Biro, W. Maurer, "Transient 3D eddy current calculations in fusion reactors", *IEEE Trans. on MAG.*, vol. 26, pp. 2364-2466, 1990.

[18] Z.J. Cendes, "Vector finite-elements for three dimensional field computation", 4th Biennial IEEE Conf. Electromagn. Field Comp., Oct. 22-24, 1990 Toronto, Ontario, Canada

[19] I. Bardi, O. Biro, K. Preis, "Finite elem scheme for 3D cavities without spurious modes", 4th Biennial IEEE Conf. Electromagn. Field Comp., Oct. 22-24, 1990 Toronto, Ontario, Canada

[20] J. Webb, "The finite-element method for finding modes of dielectric loaded cavities", *IEEE Trans. on MTT.*, vol. 33, pp. 635-638, 1985.

[21] I. Bardi, O. Biro, K. Preis, G. Vrisk, K.R. Richter, "Nodal and edge element analysis of inhomogeneously loaded 3D cavities", submitted to *COMPUMAG'91, Sorrento, 7-11 July, 1991*

A COMPARISON OF THE COULOMB AND LORENTZ GAUGE MAGNETIC VECTOR POTENTIAL FORMULATIONS FROM THE NUMERICAL POINT OF VIEW

Toshiya Morisue
Department of Chemical Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01, Japan

Abstract-In using the magnetic vector potential for eddy current calculations, there exists the problem of selecting the gauge for it. As typical gauges, there are the Coulomb and Lorentz gauges. As is well known, by the gauge transformation the former transforms to the latter, and vice versa[1].
In this paper, the gauge transformation between the Coulomb and Lorentz gauges is investigated numerically for 2D and 3D eddy current problems, using the finite difference method for the conductor region and the boundary integral equation method for the free space region.
It is concluded from the computed results that the two formulations yield almost the same results provided that an appropriate mesh is used for the conductor region.

INTRODUCTION

As is well known, the electric field intensity is expressed in terms of the magnetic vector potential and electric scalar potential as:

$$\underline{E} = -\partial\underline{A}/\partial t - \nabla V \tag{1}$$

Physical quantities such as the electric field intensity and current density remain unchanged under the following gauge transformation[1]:

$$\underline{A}' = \underline{A} + \nabla\xi, \quad V' = V - \partial\xi/\partial t \tag{2}$$

where ξ is an arbitrary scalar function. It follows as a consequence that there exist infinitely many formulations for the eddy current calculation which use the magnetic vector potential.
For example, let us assume that $\underline{A}$ satisfies the Coulomb gauge ($\nabla\cdot\underline{A} = 0$). Then, if we choose the scalar function ξ as:

$$\nabla^2\xi - \mu\sigma\partial\xi/\partial t + \mu\sigma V = 0 \tag{3}$$

the corresponding $\underline{A}'$ and V' satisfy the Lorentz gauge:

$$\nabla\cdot\underline{A}' + \mu\sigma V' = 0 \tag{4}$$

In this paper, the two typical gauges, namely the Coulomb and Lorentz gauges, are investigated numerically. In the numerical computation, the finite difference method is used for the conductor region and the boundary integral equation method for the free space region, and the Gauss-Seidel method is used as the solution method.

FORMULATION USING THE COULOMB GAUGE

For simplicity, we assume that the permeability and conductivity are constants in each region. The formulation is written as[2]:

In the conductor (Ω_1):

$$\nabla^2\underline{A}_1 - \mu\sigma\partial\underline{A}_1/\partial t - \mu\sigma\nabla V = 0, \quad \nabla^2 V = 0 \tag{5}$$

In free space (Ω_2):

$$\nabla^2\underline{A}_2 + \mu_0\underline{J}_0 = 0 \tag{6}$$

On the interface (Γ):

$$\underline{A}_1 = \underline{A}_2, \quad \underline{n}\times 1/\mu\,\nabla\times\underline{A}_1 = \underline{n}\times 1/\mu_0\,\nabla\times\underline{A}_2$$
$$\nabla\cdot\underline{A}_1 = \nabla\cdot\underline{A}_2, \quad \underline{n}\cdot[\partial\underline{A}_1/\partial t + \nabla V] = 0 \tag{7}$$

Boundary condition at infinity:

$$\underline{A}_2(\underline{r},t) = O(1/|\underline{r}|^2) \tag{8}$$

Initial condition at $t = 0$:

$$\underline{A}_1 = 0, \quad V = 0, \quad \underline{A}_2 = 0 \tag{9}$$

FORMULATION USING THE LORENTZ GAUGE

As is well known, the Lorentz gauge magnetic vector potential and electric scalar potential are written in the conductor as[3]:

$$\nabla^2\underline{A}_1 - \mu\sigma\partial\underline{A}_1/\partial t = 0, \quad \nabla^2 V - \mu\sigma\partial V/\partial t = 0 \tag{10}$$

Applying the following gauge transformation:

$$\nabla^2\xi - \mu\sigma\partial\xi/\partial t = 0 \quad \text{in } \Omega_1, \text{ and}$$
$$\partial\xi/\partial n = \int_0^t \partial V/\partial n\, d\tau \quad \text{on } \Gamma \tag{11}$$

to Eq.(10) gives:

$$\nabla^2\underline{A}_1' - \mu\sigma\partial\underline{A}_1'/\partial t = 0 \quad \text{in } \Omega_1 \tag{12}$$

$$\nabla^2 V' - \mu\sigma\partial V'/\partial t = 0 \quad \text{in } \Omega_1, \text{ and}$$
$$\partial V'/\partial n = 0 \quad \text{on } \Gamma, \quad V'(\underline{r},t) = 0 \quad \text{at } t = 0 \tag{13}$$

From Eq.(13), we obtain $V' = 0$ in Ω_1, therefore, $\underline{A}_1'$ reduces to the modified magnetic vector potential.
In the following formulation, we use the modified vector potential in the conductor and the magnetic scalar potential in free space.

In the conductor (Ω_1):

$$\nabla^2\underline{A} - \mu\sigma\partial\underline{A}/\partial t = 0 \tag{14}$$

In free space (Ω_2):

$$\underline{H} = \underline{H}_0 - \nabla\phi, \quad \nabla^2\phi = 0, \text{ and}$$
$$\underline{H}_0(\underline{r},t) = \int_{\Omega_2} \underline{J}_0(\underline{r}')\times(\underline{r}-\underline{r}')/4\pi|\underline{r}-\underline{r}'|^3\, d\Omega_2 \tag{15}$$

On the interface (Γ):

$$\underline{n}\cdot\underline{A} = 0, \quad \underline{n}\times 1/\mu\,\nabla\times\underline{A} = \underline{n}\times(\underline{H}_0 - \nabla\phi), \text{ and}$$
$$\underline{n}\cdot\nabla\times\underline{A} = \underline{n}\cdot\mu_0(\underline{H}_0 - \nabla\phi) \tag{16}$$

Boundary condition at infinity:

$$\phi(\underline{r},t) = O(1/|\underline{r}|^2) \tag{17}$$

Initial condition at $t = 0$:

$$\underline{A} = 0, \quad \phi = 0 \tag{18}$$

NUMERICAL COMPARISON OF THE FORMULATIONS

We consider a 3D time-harmonic eddy current problem for an aluminum block shown in Fig.1, placed in a uniform magnetic field.
Physical properties used are as follows:

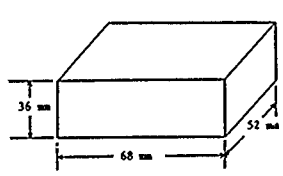


Fig.1 Aluminum Block

$\mu = 4\pi \times 10^{-7}$ H/m; $\sigma = 2.5 \times 10^{7}$ S/m,

$f = 50$ and 100 Hz, $B_0 = 1 + j0$ T (19)

The discretization of the block is shown in Fig.2. The total number of nodes for the finite difference equations is 2,368 and that for the boundary integral equations is 832. The computation is carried out for the 1/8 geometry, using a symmetry relation. The shape of the element for FDM is a cube whose size is 4 mm x 4 mm x 4 mm, and that for BIEM is a square whose size is 4 mm x 4 mm. Zero-order interpolation is used for BIEM. The solution method used is the Gauss-Seidel method.

Computed results for the eddy current distribution in the block are shown in Fig.3 for the frequency of 50 Hz and in Fig.4 for the frequency of 100 Hz. Computer used is HP9000-360. The computation times for the Coulomb gauge and Lorentz gauge formulations are respectively 74 min 55 sec and 57 min 54 sec.
It is seen in Fig.3 and Fig.4 that there exists a discrepancy of 10~15 percent between the two formulations. The cause for the discrepancy is low discretization of the block. The discrepancy becomes smaller with higher discretization. This fact has already been proved for a 2D case.

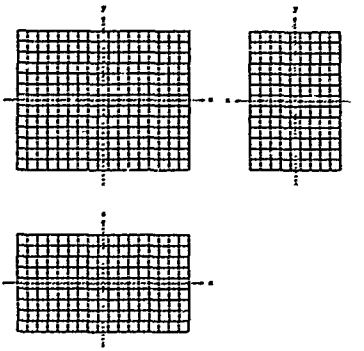


Fig.2 Discretization of The Block

CONCLUSION

In this paper, the Coulomb and Lorentz gauge magnetic vector potential formulations for 3D eddy current calculations are investigated numerically for a 3D eddy current problem. From the computed results, it may be concluded that the two formulations yield almost the same results provided that an appropriate discretization is applied to the conductor region.
For the problem having a constant conductivity, the Lorentz gauge formulation has an advantage over the Coulomb gauge formulation in the computation time. However this advantage fails for the problem having a changing conductivity.
This work was partially supported by Grant-in-Aid for Scientific Research 01302031 from the Ministry of Education, Science and Culture in Japan

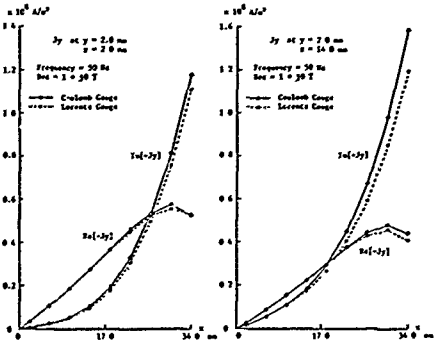


Fig.3 Computed Eddy Currents, Frequency = 50 Hz

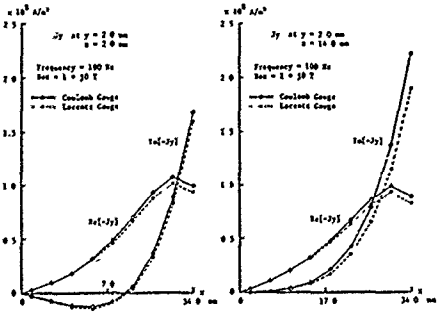


Fig.4 Computed Eddy Currents, Frequency = 100 Hz

REFERENCES

[1] J.D.Jackson, Classical Electrodynamics. John Wiley & Sons, 1975, ch.6, pp.220-223
[2] T.Morisue, "Three-Dimensional Eddy Current Calculation in Multiply Connected Regions by Using the Magnetic Vector Potential," International Journal for Numerical Methods in Fluids, Vol.11,No.6, 1990, pp.881-891
[3] J.C.Slater and N.H.Frank, Electromagnetism. Dover Publications, 1969, ch.7, pp.86-88

HYBRID MAGNETIC FORMULATION FOR MULTIPLY CONNECTED PROBLEMS

Z. Ren and F. Bouillault

Laboratoire de Génie Electrique de Paris, U.R.A.127 CNRS, ESE, Univ.Paris 6 & 11
Plateau du Moulon, 91192 Gif sur Yvette cedex, France

Introdiction

In electromagnetic field computation, the magnetic scalar potential is usually employed in the air space. However, the use of a scalar potential brings the multi-valued problem when the studied domain is multiply-connected. The cutting surfaces or the cutting lines should be introduced for solving this problem [1][2].

It has been shown that, in a hybrid finite element - boundary integral formulation in terms of electric field [3], to avoid the multi-valued problem due to the scalar potential, one can take the electric field e as working variable in the conducting region as well as in the air space. The electric field in the air space is expressed by a boundary surface current distribution. The use of a spanning tree technique [4] permits to ensure the divergence free of the surface current density.

In this paper, we show that the multiply connected problem can also be solved in a magnetic formulation, provided that the vector variable h is used as working variable instead of a scalar potential in the air space. The magnetic field h in the air can be expressed by a boundary surface current k as we do in the electric formulation. The vector variables h and k on the boundary are discretised by boundary edge elements. The employement of the boundary tree technique [4] ensures the curl-free of h and the div-free of k. The number of boundary unknowns is associated with the number of branches of the tree which has the same order as the boundary nodes. Comparing with the eletric formulation, the number of boundary unkowns is less important.

Hybrid Magnetic Formulation

Solving weakly the Faraday's law $curl\, \mathbf{e} = -\partial \mathbf{b}/\partial t$, we get the following variational formulation in terms of magnetic field h :

$$\int_\Omega \frac{1}{\sigma} curl\, \mathbf{h}' \cdot curl\, \mathbf{h}\, d\Omega + \frac{d}{dt}\int_\Omega \mu\, \mathbf{h}' \cdot \mathbf{h}\, d\Omega + \int_\Gamma \mathbf{h}' \cdot \mathbf{n} \times \mathbf{e}_r\, d\Gamma = -\int_\Gamma \mathbf{h}' \cdot \mathbf{n} \times \mathbf{e}_0\, d\Gamma \qquad (1)$$

where Ω is the conducting region bounded by Γ, σ and μ are the conductivity and the permeability of the conductor, h' is a test function. In exterior region, the electric field e is the sum of a reduced field $\mathbf{e}_r$ and a source field $\mathbf{e}_0$ due to the excitation current.

The volume integral of eq.(1) is discretised by the tetrahedral edges elements as in [5].

The exterior region of Ω is taken into account by the surface integral term on the boundary Γ. Instead of introducing a magnetic scalar potential in the air space [5], we use the reduced magnetic field $\mathbf{h}_r$ as unknown variable. Eq.(1) can be solved if a relationship is established between the tangential component of $\mathbf{e}_r$ and the tangential component of $\mathbf{h}_r$.

Introducing a surface current distribution k on Γ, $\mathbf{n} \times \mathbf{e}_r$ and $\mathbf{n} \times \mathbf{h}_r$ at an any point x on Γ are given by

$$\mathbf{n}_x \times \mathbf{e}_r(\mathbf{x}) = -\frac{\mu_0}{4\pi}\frac{d}{dt}\int_\Gamma \frac{\mathbf{n}_x \times \mathbf{k}(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\, d\Gamma \qquad (2)$$

and

$$\mathbf{n}_x \times \mathbf{h}_r(\mathbf{x}) = \frac{1}{2}\mathbf{k}(\mathbf{x}) - \frac{1}{4\pi}\int_\Gamma \frac{\mathbf{n}_x \times ((\mathbf{x}-\mathbf{y}) \times \mathbf{k}(\mathbf{y}))}{|\mathbf{x}-\mathbf{y}|^3}\, d\Gamma \qquad (3)$$

where $\mathbf{n}_x$ is the outward unit normal at the boundary point x and $|\mathbf{x}-\mathbf{y}|$ is the distance between x and a boundary point y.

Eqs.(2) and (3) are discretised by a variational method by multiplying test functions h' to (2) and k' to (3). Elimilating the surface current density k from (2) and (3) after the boundary discretisation, we relate $\mathbf{e}_r$ and $\mathbf{h}_r$ by a matrix equation.

Substituting $\mathbf{e}_r$ by $\mathbf{h}_r$ in the finite element matrix equation obtained from eq.(1), we get the whole matrix system. The degrees of freedom are the circulation of the magnetic field on the inner edges of the conductor and on the independent boundary edges determined by the spanning tree technique.

Boundary Element Discretisation

The boundary of the studied domain is meshed by triangles. The reduced magnetic field on the boundary is approximated by using the triangular edge elements preserving the tangential continuity. In a triangle, we have

$$\mathbf{h}_r = \sum_{m=1}^{3} \mathbf{w}_m h_m \qquad (4)$$

where the unknown coefficient h_m is the circulation of $\mathbf{h}_r$ along the edge m. $h_m = \int_m \mathbf{h}_r \cdot \mathbf{l}\, dl$ and $\mathbf{w}_m$ is a vector interpolation function defined by the barycenter coordinates λ_Γ:

$$\mathbf{w}_m = \lambda_{\Gamma i}\, grad\lambda_{\Gamma j} - \lambda_{\Gamma j}\, grad\lambda_{\Gamma i} \qquad (5)$$

with i, j the two extremities of the edge m.

Similarly, the surface current density in a triangle is interplolated by

$$\mathbf{k} = \sum_{m=1}^{3} \mathbf{v}_m k_{r_3} \qquad (6)$$

where the unknown coefficient k_m is the line integral of k along the edge m: $k_m = \int_m \mathbf{k} \cdot \mathbf{s}_m dl$ with $\mathbf{s}_m$ the unit vector in the plane of the triangle normal to the edge m, and

$$\mathbf{v}_m = \mathbf{n} \times (\lambda_{ri} grad\lambda_{rj} - \lambda_{rj} grad\lambda_{ri}) \qquad (7)$$

The coefficient k_m represents in fact the surface current across the edge m. The so defined triangular edge element ensures the normal continuity of k on a common edge of two adjacent triangles if the normal vector $\mathbf{s}_m$ is well defined.

The previously described triangular edge elements provide the tangential continuity of h or the normal continuity of k, but not the curl-free of h or the div-free of k.

The curl-free of h and the div-free of k on the boundary surface is equivalent to the zero curvilinear integrals on a closed simply connected path (the closed path do not link the conductor and do not around the hole of the multiply connected domain):

$$\oint \mathbf{h} \cdot \mathbf{l} dl = \sum \int_m \mathbf{h}_r \cdot \mathbf{l} dl = \sum h_m = 0$$

$$\oint \mathbf{k} \cdot \mathbf{s} dl = \sum \int_m \mathbf{k} \cdot \mathbf{s}_m dl = \sum k_m = 0$$

The above expressions mean that, on the triangular mesh, only a set of edges is independent.

It can be shown that the number of independent edges is the number of branches of a tree plus two times the number of holes of a multiply connected region. An automatical tree generation algorithm is developped to identify the independant edges [4].

Example and Discussion

The following example is a benchmark problem of TEAM Workshop [6]. It concerns the analysis of eddy currents in a asymmetrical aluminum plate with a hole when the excitation coil is supplied with a sinusoidal current.

We solve this problem by the described hybrid formulation in terms of magnetic field h. The conductor is meshed by 2367 tetrahedra with 3471 edges. The boudary mesh has 452 nodes, 906 triangles and 1356 edges. The number of degrees of freedom on the boundary of the domain is the number of independent edges which is equal to 453.

The same problem with the same mesh has been solved by the electric formulation [7], where the boundary unknowns are associated with the boundary edges which is equal to 1356. The boundary matrix calculation is more consumed in this case.

The z-component flux density along a line ($y = 72mm, z = 34mm$) obtained by two dual methods (magnetic and electric formulations) are compared with measured ones in the Fig.1. A good agreement between the calculated and measured values is observed.

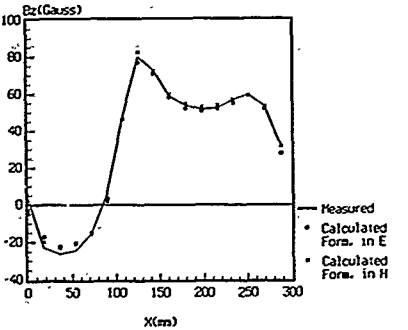


Fig.1. z-component flux density (real part) along the line $y = 72mm$, $z = 34mm$.

Conclusion

In a hybrid magnetic formulation, the multiply connected problems can be solved by using the magnetic field as unknown variable in both the conducting region and the air space. The use of the boundary spanning tree technique permits to ensured the curl free of the magnetic field and leads to a reduction of boundary unknowns. The numerical example shows the validity of the method.

References

[1] C.S.Harrod and J.Simkin, 'Cutting multiply connected domains', IEEE Trans., 1985, Mag-21, (6), pp. 2495-2498

[2] J.C.Vérité, 'Calculation of multivalued potentials in exterior regions', IEEE Trans., 1987, Mag-23, (3), pp. 1881-1887

[3] Z.Ren, F.Bouillault, A.Razek, A.Bossavit and J.C.Vérité, 'A new hybrid model using electric field formulation for 3-D eddy current problems', IEEE Trans., 1990, Mag-26,(2), pp.470-473

[4] Z.Ren and A.Razek, 'New technique for solving 3-D multiply connected eddy current problems', IEE proceedings, 1990, Vol.137, Pt.A, No.3, pp.135-140

[5] A.Bossavit and J.C.Vérité, 'The "Trifou" code. solving the 3D eddy-currents problem by using h as state variable', IEEE Trans., 1983, Mag-19, (6), pp. 2465-2470

[6] L.R.Turner, K.Davey, C.R.I.Emson, K.Miya, T Nakata and A Nicolas, 'Problems and workshops for eddy current code comparison', IEEE Trans., 1988, MAG- 24, (1), pp.431-434

[7] Z.Ren and A.Razek, 'Calculation of 3-D eddy currents using electric formulation', Proceedings of TEAM Workshop, Bièvre,France,March,1989

OPTIMIZATION OF SHIELDING SYSTEMS

Konrad Weeber[1], S. Ratnajeevan H. Hoole
Department of Engineering, Harvey Mudd College, Claremont, CA 91711, USA
1) on leave from Laboratoire d'Electrotechnique, ENSIEG, 38402 St. Martin d'Heres, France

ABSTRACT: The analysis of electromagnetic devices under the influence of eddy currents, in particular shielding structures, is a well known and developed art as of today [1-4]. Design synthesis methods as opposed to analysis methods have recently emerged as a viable tool in the optimization of device performance in order to meet specified goals. This paper applies finite element optimization techniques and a newly developed geometric parametrization method to the shape optimization of electromagnetic shielding systems.

1)FINITE ELEMENTS AND OPTIMIZATION TECHNIQUES

In the optimal design of shielding devices we generally want to achieve a desired flux density distribution within the shielded region by appropriately varying the design parameters such as the material and geometric properties of the shield itself. For this purpose an object function F is defined, which is at its minimum when the design of the shield is optimal. For example, the object function

$$F = \sum_i (B_i - B_{opt,i})^2 \qquad (1)$$

describes the deviation of the computed flux density B, as obtained for a given trial geometry, from the desired field distribution B_{opt} at given sampling points i within the shielded region.

To approach the minimum of the object function systematically, gradient methods have been found to perform fastest [5],[6]. In the course of these powerful gradient methods, the optimum is approached in successive line searches: at each iteration the improved parameter vector p^{i+1} is found from the previous one by minimizing the object along a search direction S, performing essentially a search for the ideal value of the scalar α

$$p^{i+1} = p^i + \alpha S \qquad (2)$$

Which way the search direction itself is chosen, depends on the gradient method employed (conjugate gradient or steepest descent method), but it basically relies on the gradient of the object function F in eq. (1) with respect to all parameters p. In this essential part of the optimization procedure, the gradient dF/dp is obtained as summation of terms of the following form

$$\frac{dF}{dp} = \frac{\partial F}{\partial p} + \frac{\partial F}{\partial B}\frac{\partial B}{\partial p} = 2(B - B_{opt})\frac{\partial B}{\partial p} \qquad (3)$$

Here the first term of the partial derivatives generally vanishes due to the nature of the object function, which does not explicitly depend on p. The magnetic flux density B, related to the magnetic vector potential A by the differential operator "rot", is expressed in the finite element discretization within an element in terms of the first order differentiation vectors {d} and the nodal potential values {A} as $B = \{d\}^t\{A\}$, [5]. Thus the parameter sensitivity of the flux density within a finite element is

$$\frac{\partial B}{\partial p} = \frac{\partial \{d\}^t}{\partial p}\{A\} + \{d\}^t\frac{\partial \{A\}}{\partial p} \qquad (4)$$

with the first term as the sensitivity of the first order differentiation vectors and the second one as the sensitivity of the nodal values of the magnetic potential. In general we specify the desired field values B_{opt} in the shielded region, which is not part of the parameter dependent shielding structure itself, so that the first term vanishes. The gradient of the object function thus merely depends on the parameter sensitivity of the magnetic potential A.

This potential sensitivity may be evaluated directly from the finite element solution without the need for an additional field solution [7]. In the case of eddy current analysis, the complex valued finite element equation

$$\left([R(\mu,x,y)] + j[Q(\omega,\sigma,x,y)]\right)\{A\} = \{P\}\{A\} = \{J\} \qquad (5),$$

as derived from the diffusion equation for the magnetic vector potential A, is differentiated with respect to the design parameters p to yield

$$[P]\left\{\frac{\partial A}{\partial p}\right\} = \left\{\frac{\partial J(x,y)}{\partial p}\right\} - \left[\frac{\partial P(\mu,\sigma,\omega,x,y)}{\partial p}\right]\{A\} \qquad (6)$$

The coefficient matrix [P] has been previously assembled during the standard field solution of eq. (5). In applying Choleski decomposition schemes as matrix solvers, we can use this decomposed matrix [P]; the only computational effort for obtaining the many potential gradients is then just the assemblage of the right hand side and forward elimination and back substitution for each parameter [5,8].

The assemblage of the right hand side can be classified into the two cases of material optimization and shape optimization. In the first case of material properties as parameters, the partial differential of the material independent source term vanishes. The explicit dependence of the coefficient matrix on the material parameters renders straightforward partial derivatives [7]. It is the second case which is of special interest in this paper, where the optimization parameters are used to describe the shape of the geometry which we want to optimize. Then the derivatives on the right hand side of (6) have to be evaluated for each element as

$$\left\{\frac{\partial J_{el}}{\partial p}\right\} - \left[\frac{\partial P_{el}}{\partial p}\right]\{A\} = \sum_i \left\{\frac{\partial J_{el}}{\partial x_i}\right\}\frac{\partial x_i}{\partial p} - \left[\frac{\partial P_{el}}{\partial x_i}\right]\frac{\partial x_i}{\partial p}\{A\} \qquad (7)$$

with x_i as the spatial coordinates of all element nodes. The derivatives of the element matrices with respect to the locations of the vertices follow directly from the finite element matrices using the first order differentiation vectors [7]. It is the calculation of the gradient of the nodal locations that arises in shape optimization and that requires further investigation, as outlined in the following section.

2)GEOMETRY PARAMETRIZATION AND DESIGN MODEL

The successful application of gradient algorithms for the process of shape optimization is frequently jeopardized by discontinuities in the object function. As the geometry of the device changes in the iterative search for the minimum, a new finite element mesh is necessitated for each new geometric parameter value. In employing free meshing, topologically different meshes arise, each one representing a possible domain discretization with an inherent error. This change of mesh topology and discretization error occurs in discrete steps and leads to discontinuities in the object function. The now just piecewise continuous object function obstructs a smooth convergence towards the minimum [9]. Additionally, we have to acknowledge that especially in eddy current analysis sharp changes of field strengths within the often relatively small penetration depths have to be modeled, so that any finite element mesh is an approximation, often relatively crude. This holds particularly for 3D analyses, where one knowingly accepts a larger approximation error in order to keep the computational effort affordable [3].

The basic requirements for the successful application of gradient methods, however, are that the object function gradients, the mesh distortions, and the finite element discretization errors vary sufficiently smoothly. As the need for parametrization is rather to describe a variation to a given geometry than to define it, it has been shown [10] that all requirements are ideally fulfilled, when the finite element mesh for new geometric parameter values is obtained through a mapping technique, which is based on similarities with optimization problems in civil engineering [11]. In the course of applying this method, we enclose the area to be parametrized and some surrounding domain within a subregion, for which we perform a structural finite element analysis [12],[13]. The solution of the displacement vector {u} resulting from applied forces {F} to a structure described by the stiffness matrix [K] is obtained from

$$[K]\{u\} = \{F\} \qquad (8)$$

With this solution for the displacement {u} the spatial coordinates {x} of the structure deflected from its initial shape $\{x_0\}$ are given as

$$\{x\} = \{x_0\} + \{u\} \qquad (9)$$

If we do not "excite" the structural displacements by applied forces to the structure, but rather by specified displacements u_s at the surfaces of the design geometry, we obtain the unknown displacements by appropriate partitioning of the stiffness matrix as solution to

$$[K_{uu}]\{u_u\} = [K_{us}]\{u_s\} \qquad (10)$$

The applied surface displacements $\{u_s\}$ are now employed as optimization parameters p. The solution of eq. (10) is then basically the mapping of the parameters as surface displacements u_s onto the whole structural solution domain, giving the structural deformation and hence the smooth deformation of the finite element mesh for each new parameter value p. The solution of eq. (10) renders a displacement space, from which we can deflect and distort our finite element mesh during the parameter changes of a line search by linear superposition, replacing the need for free meshing. In addition, if we choose unit applied surface displacements u_s, the sensitivities of the nodal coordinates are also given as the solution of eq. (10) itself

$$\frac{\partial x_i}{\partial p} = \frac{\partial (x_0+u)}{\partial p} = \frac{\Delta u}{\Delta p} = \Delta u = u_u \qquad (11)$$

Thus, for the cost of an additional structural solution per line search with a model containing just the regions of variable geometry, the required smooth changing discretization error is achieved. Furthermore, smooth shape outlines are obtained even with only few optimization parameters, as will be shown in the following example.

3) AN EXAMPLE : SHAPE OPTIMIZATION OF A SHIELD

For the purpose of demonstrating the methodologies described in the previous sections, we investigate the shape optimization of an electromagnetic shield. Due to the eddy currents induced in the shield, the magnetic field is prevented from penetrating into the region behind the shield. In this example, the shield itself is interrupted by slots, like cooling slots in the housing of an electric device. The field, as originating from the source coil on the left side, shall not exceed, or in the case of minimal shield volume, be exactly the desired flux density of 0.01 T at the right side of the shield (Fig. 1). Without using a shielding at all, the flux densities along the line of object evaluation are in the range of 0.12 to 0.15 T. As initial guess for the shape of the shield a rectangular crossection is chosen, with a thickness of twice the penetration depth and a height equal to the distance, along which the desired flux density is specified. 7 parameters are used to modify this shape, where the 7th one is the uniform displacement of the top shield face. The location of the right shield face and the location and width of the slots are fixed. The optimum shape of Fig. 2 has been reached within only 14 conjugate gradient line searches without the problem of encountering discontinuities in the object function. As expected, the shield crossection narrows towards the top, where the need for shielding reduces. Above the slot at the line of symmetry, the shield crossection approaches a thickness of zero, so that this slot may be manufactured in fact wider without violation of the object function. The flux density within the shielded region approaches the desired value closely (Fig. 3).

4) SUMMARY

A robust and reliable parametrization method has been presented for the shape optimization of eddy current devices. The discontinuities in the object function, due to the finite element discretization, are eliminated by a structural mapping technique, which is especially useful for geometries of high complexity.

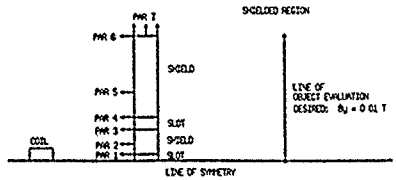


Fig. 1) Problem definition and starting geometry of the shield

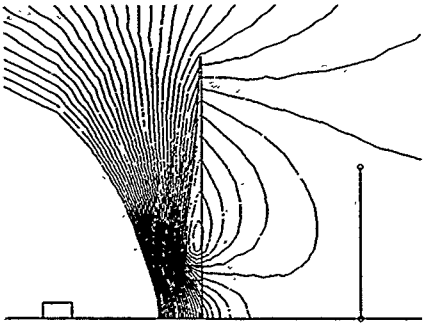


Fig. 2) Shield geometry and contour lines of the magnetic potential A at the optimum shield design (contour lines displayed only in the vicinity of the shield)

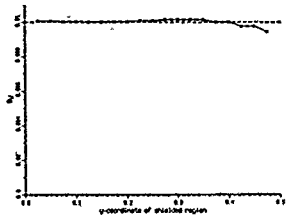


Fig. 3) Flux density distribution along line of object evaluation

LIST OF REFERENCES :

[1] Richard L. Stoll, "The Analysis of Eddy Currents", Clarendon Press, Oxford 1974

[2] M.V.K. Chari, "Finite Element Solution of the Eddy Current Problem in Magnetic Structures", IEEE Trans Power Apparatus and Systems, Vol.PAS-93, No.1, Jan 1974

[3] T. Nakata, N. Takahashi, K. Fujiwara, T. Imai, K.Muramatasu, "Comparison of Various Methods of Analysis and Finite Elements in 3D Magnetic Field Analysis", 4th IEEE Conf on Electromagnetic Field Computation, Oct. 1990, Toronto, Canada

[4] S.R.H. Hoole, "Computer-Aided Analysis and Design of Electromagnetic Devices", Elsevier, 1989

[5] S. Gitosusastro, J.L. Coulomb, J.C. Sabonnadiere, "Performance Derivative Calculations and Optimization Process", IEEE Trans. Magn., Vol 25, No.4, July 1989

[6] G.Vanderplaats, "Numerical Optimization Techniques for Engineering Design", McGraw-Hill, 1933

[7] S.R.H. Hoole, "Inverse Problems - Finite Elements in Hop-Stepping to speed up", International Journal of Applied Electromagnetics in Materials,1,1990, p. 255-261

[8] S.R.H. Hoole, "Optimal Design, Inverse Problems and Parallel Computers", 4th IEEE Conf. on Electromagnetic Field Computation, Oct. 1990, Toronto, Canada

[9] S. Subramaniam, K. Weeber, S.R.H. Hoole, "Smoothness of Object Functions, Finite Element Meshes and Edge Elements in Electromagnetic Device Synthesis", 5th MMM-Intermag Conf., June 1991, Pittsburgh, Pennsylvania

[10] K. Weeber, S R.H. Hoole, "A Structural Mapping Technique for Geometric Parametrization in the Synthesis of Magnetic Devices", Int J.Num.Meth.Eng., in review process

[11] A.Irrgang, I Raasch, "Shape Optimization with MSC/NASTRAN", MacNeal-Schwendler European Users' Conference, 1988

[12] O C.Zienkiewicz, R.L.Taylor, "The Finite Element Method", McGraw-Hill, 4th. ed., 1988

[13] W.Weaver, P.R.Johnston, "Finite Elements for Structural Analysis", Prentice Hall, 1984

3D Eddy Current Computation

Nathan Ida
Department of Electrical Engineering
The University of Akron
Akron, Ohio 44325-3904, U.S.A

Abstract - Two different formulations for computation of eddy current fields are presented. These are representative of methods that use potential functions on the one hand and field variable on the other. The two methods also contrast the finite element method and the coupled finite element, boundary element methods. In terms of approximations, one uses standard nodal-based elements while the other the vector (edge) elements. Results typical to those obtainable from these formulations are shown.

INTRODUCTION

The computation of 3D eddy currents is normally done by one of two basic methods. The first defines potential functions to represent the magnetic field[1]. The most common method, is the use of the magnetic vector potential A (defined as $B=\nabla XA$). To this, a scalar function is often added. The divergence condition on the vector A is specified through the Coulomb or Lorenz gauges. In essence, this method is a direct extension of the two dimensional formulations where the use of the magnetic vector potential is almost universal. The advantage of using A in the two dimensional case is obvious: since A has only one component, it is considered to be a scalar, resulting in a single degree of freedom per node in the finite element mesh. In the 3D case, three components must be used to which, in most cases, a scalar potential must be added, resulting in four degrees of freedom per node. These formulations use almost exclusively, first or second order scalar (nodal-based) finite elements.

An alternative method is to use the field variables B or E directly[2]. In this form, Maxwell's equations are written in terms of one of the variables and these are approximated over a finite element mesh. Because of the continuity requirements for the field variables, vector finite elements are more attractive than scalar elements. These elements guarantee tangential continuity.

Two formulations, one a finite element method based on the magnetic vector potential and one a coupled finite element-boundary integral formulation based on E (or H), as representative of the many formulations that exist, are outlined here. The first uses a first order (brick) nodal-based finite element while the second uses a first order (tetrahedral) vector (edge) element and a corresponding triangular edge element on the surface.

FORMULATIONS

Using the magnetic vector potential A, the curl-curl equation representing eddy current phenomena is

$$\frac{1}{\mu}\nabla\times\nabla\times A = J_s + J_e \tag{1}$$

where J_s is the source current density and J_e is the induced eddy current density. Displacement currents are normally deleted from eddy current formulations. The eddy currents are:

$$J_e = -\sigma\left(\frac{\partial A}{\partial t} + \nabla V\right) \tag{2}$$

where V is the electric scalar potential. In eddy current problems, because eddy currents are not arbitrary, it is important to restrict the eddy currents in the solution region. This can be done by introducing a constraint equation. A simple way to constrain the eddy currents is to use the current continuity equation $\nabla J_e=0$ as a constraint. The field equation together with the constraint equation are

$$\frac{1}{\mu}\nabla\times\nabla\times A = J_s - \sigma\left(\frac{\partial A}{\partial t} + \nabla V\right) \qquad \nabla\cdot\left(\sigma\frac{\partial A}{\partial t} + \sigma\nabla V\right) = 0 \tag{3}$$

The divergence of A is enforced using Coulomb's gauge ($\nabla.A=0$). With this equation (3) becomes

$$-\frac{1}{\mu}\nabla^2 A = J_s - \sigma\left(\frac{\partial A}{\partial t} + \nabla V\right) \qquad \nabla\cdot\left(\sigma\frac{\partial A}{\partial t} + \sigma\nabla V\right) = 0 \tag{4}$$

Equation (4) is the governing equation for the constrained A-V formulation.

This formulation can be modified to account for other phenomena. First, since displacement currents have been deleted, the equation in (4) cannot account for propagation aspects of fields. By adding the displacement currents back into the formulation, the basic equation in (4) can be used to compute resonant frequencies in cavities as well as losses[3]. The formulation now looks as:

$$\nabla^2 A = \mu J_s - \omega^2\mu\varepsilon A - \sigma\mu(j\omega A + \nabla V) \qquad \nabla\cdot(j\sigma\omega A + \sigma\nabla V) = 0 \tag{5}$$

Another modification consists of adding a velocity term to equation (4). This yields an eddy current formulation with velocity effects that allows the computation of fields of moving devices[4].

In considering the inclusion of velocity terms in the eddy current equations, the total electric field is: $E=uX\nabla XA-j\omega A-\nabla V$, where u is the velocity vector. The governing equations with the velocity term are:

$$\nabla\times\frac{1}{\mu}\nabla\times A - \nabla\frac{1}{\mu}\nabla.A = J_s + \sigma u\times\nabla\times A - j\omega\sigma A - \sigma\nabla V \tag{6a}$$

$$\nabla.\left[\sigma u\times\nabla\times A - j\omega\sigma A - \sigma\nabla V\right] = 0 \tag{6b}$$

The solution of these equations (4), (5), or (6) leads to the correct field quantities based on the general field representation at any frequency. The magnetic flux density is calculated from ∇XA and the electric field from $E=-(j\omega A+\nabla V)$ or $E=(uX\nabla XA-j\omega A-\nabla V)$. Other quantities, such as the Q-factor of a lossy cavity can be computed from the fields.

A second method, based on the field variable E (or alternatively, H) is to solve:

$$\nabla\times E = -j\omega\mu H \qquad \nabla\times H = j\omega\varepsilon' E \qquad \text{where } \varepsilon'=\varepsilon+\sigma/j\omega \tag{7}$$

or, by eliminating H, we get for E:

$$\nabla\times\frac{1}{j\omega\mu}\nabla\times E + j\omega\varepsilon E = 0 \tag{8}$$

In this form, both eddy currents and displacement currents are taken into account through the complex permittivity ε'. However, since a coupled method is sought we assume that the source is external to the solution domain. A weak form of this may be written as

$$\int_v \left(\frac{1}{j\omega\mu}\nabla\times\mathbf{E}\right)\cdot(\nabla\times\mathbf{w}_m)dv + \int_v j\omega\varepsilon\mathbf{E}\cdot\mathbf{w}_m dv = \int_s (\hat{n}\times\mathbf{H})\cdot\mathbf{w}_m ds \qquad (9)$$

where $\mathbf{w}_m$ are a set of vector weighting functions. These weighting functions are the vector element shape functions[5] and, together with the Galerkin method for discretization of the weak form provide a method for solving the eddy current problem. The current density on the surface $J=\hat{n}\times H$, is used as a nonlocal boundary condition to couple the boundary integral method on the surface and the finite element method in the volume. Using the approximations $E=\Sigma E_k w_k$ and $\hat{n}\times H=\Sigma(\hat{n}\times H)_k w_k=\Sigma F_k w_k$ for an element k in the volume, and $E=\Sigma E_l f_l$ and $\hat{n}\times H=\Sigma(\hat{n}\times H)_l f_l=\Sigma F_l f_l$ for a boundary element l on the surface of the solution domain, provides two coupled equations:

$$[A]\{E\} = [B]\{F\} \qquad \text{for the volume mesh} \qquad (10)$$

$$[C]\{F\} = -\{E^i\} + [D]\{E\} \qquad \text{for the surface mesh} \qquad (11)$$

where $\{E^i\}$ is the incident electric field on the surface of the solution domain. Solution of this coupled system provides a method for an coupled FEM-BEM solution.

RESULTS

As examples to the use of these two formulations, two problems are solved. The first consists of an aluminum bar 2.54cm X 2.54cm X 20.32cm, with a coil, 8.128cm long and 0.685cm thick, at the center of the bar. The bar has conductivity of $3.54X10^7$S/m. The losses due to eddy currents in the bar are calculated. The results, shown in table 1 are obtained using equation (3) and compared with experimental data given in [6].

Table 1. Losses in the aluminum bar for different currents in the coil.

Current	1A	1.5A	2.0A
Numerical	0.593	1.334	2.372
Experimental	0.58	1.56	3.52

While the results for 1A agree well, the results for higher currents do not. This is due to the heating during the experiment, as was already pointed out in [6].

The second problem is shown in figure 1. A perfectly conducting cube, of electric size ks=2, where s is the side width, is placed in an incident field with the electric field in the z direction, propagating in the +y direction as shown. The looping surface current on the surface along the line abcd is shown. This particular problem was solved for two reasons: one is because comparison data is available in [7] and the second, because it demonstrates the capability of the method to calculate scattering problems. The same problem can be solved with a finite conductivity, or with a lossy dielectric replacing the conductor. The only difference is the value of the complex dielectric ε'

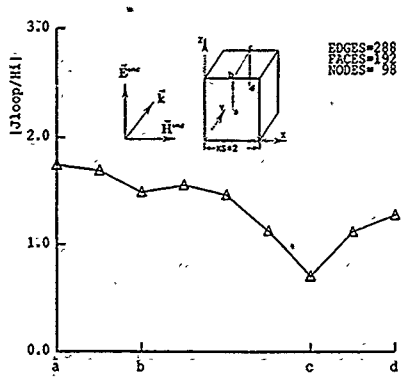


Figure 1. Magnitude of looping current in a conducting cube.

CONCLUSIONS

The formulations and results presented here are representative of eddy current problems and their formulation using finite elements. While no particular method can be claimed to be better than the others, each has its own advantages. However, the use of vector shape functions seems to be more efficient and better suited for computation of vector fields.

REFERENCES

[1] O. Biro and K. Preis, "On the Use of the Magnetic Vector Potential in the Finite Element Analysis of Three-Dimensional Eddy Currents", IEEE Transactions on Magnetics, Vol. MAG-25, No. 4, pp 3145-3159, 1989.

[2] J.S. van Welij, "Computation of Electromagnetic Fields in Terms of H on Hexahedra", IEEE Transactions on Magnetics, Vol. MAG-21, No. 6, 1985, pp. 2239-2241.

[3] J.R. Brauer, R.H. Vander Heiden and A.B. Bruno, "Finite Element Modeling of Electromagnetic Resonators and Absorbers," Journal of Applied Physics, Vol. 63, No 8, April 15, 1988, pp. 3197-3199.

[4] S. Hong and N. Ida, "Modeling of Velocity Terms in 3D Eddy Current Problems", Submitted for Presentation at Compumag, Italy, July 7-11, 1991.

[5] M.L. Barton and Z.J Cendes, "New Vector Finite Elements for Three-Dimensional Magnetic Field Computation", Journal of Applied Physics, Vol. 61, No. 8, 1987, pp. 3919-3921.

[6] N. A. Demerdash, O. A. Mohammed, T. W. Nehl, F. A. Fouad, R. H. Miller, "Solution of Eddy Current Problems Using Three Dimensional Finite Element Complex Magnetic Vector Potential", IEEE Trans. on Power App. and Syst., Vol. PAS-101, No. 11, pp. 4222-4229, 1982.

[7] K. Umanshakar, A. Taflove and S.A. Rao, "Electromagnetic Scattering by Arbitrary Shaped Three-Dimensional Homogeneous Lossy Dielectric Objects", IEEE Transactions on Antennas and Propagation, Vol. AP-34, No. 6, 1986, pp. 758-766.

MODELLING LOW FREQUENCY ELECTROMAGNETIC COMPATIBILITY USING 3D FINITE ELEMENTS

D. Rodger, T. Seddon and P.J. Leonard
School of Electronic & Electrical Engineering
Bath University
BATH, BA2 7AY, U.K.

Abstract

The suitability of a 3D finite element code (MEGA) for modelling electromagnetic compatibility problems (EMC), in the frequency range up to 40 kHz is assessed. Results from a simple shielding experiment are compared with predictions.

Introduction

The problem of shielding electrical equipment from electromagnetic interference has always been important, but new urgency has been introduced by the need to comply with a recent EEC directive 82/499/EEC which states that 'A permit is required to operate any equipment or installation producing electromagnetic oscillations from 10 kHz to 3000 GHz'.

Switching frequencies of 20 - 40 kHz have become common in power electronics equipment. A method for modelling the shielding for such apparatus is obviously important. This paper describes work done to validate the use of a 3D finite element electromagnetics program (MEGA) for modelling shielded equipment in this frequency range. An experiment consisting of a coil shielded by copper sheets arranged on the faces of a cube was devised and measurements were taken of fields outside the cube at various frequencies from 5 to 40 kHz.

Theory

The method described here [1] allows nonmagnetic conducting sheets which are thin compared to their other dimensions and of arbitrary 3D shape to be modelled economically, the use of standard volume finite elements to model the sheet is of course possible but very uneconomic if the usual requirements for aspect ratio of elements is to be met.

Non conducting regions surrounding the sheets are modelled using total or reduced magnetic scalar potentials ψ, so a Laplacian type equation of the form div μ grad ψ arises (from div B = 0).

If eddy currents flow in a conductor which is thin compared to the skin depth of the material, the distribution of current can be taken to be uniform across the sheet. If we define a surface current K as K = Jd, where d is the thickness of the sheet then, as K varies only along the sheet surface, K can be represented by means of a scalar quantity T which exists only on the sheet surface:

$$K = \frac{\partial}{\partial t} (\text{grad } T \times n) \qquad (1)$$

In the above, n is the normal to the sheet and the $\frac{\partial}{\partial t}$ term is introduced in order to achieve symmetry in the final matrix.

Figure 1 shows a typical sheet conductor, surrounded by the magnetic scalar potention ψ. If the potentials on either side of the sheet are labelled ψ_1 and ψ_2, at an arbrary point on the sheet, since $-H_1 \times n + H_2 \times n = K$, then:

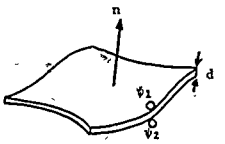


figure 1 A typical thin sheet

$$\text{grad } \psi_1 \times n - \text{grad } \psi_2 \times n = \frac{\partial}{\partial t} (\text{grad } T \times n) \qquad (2)$$

dividing by σd and taking the curl of equ (2) yields:

$$\text{curl } \frac{1}{\sigma d} (\text{grad } \psi_1 \times n - \text{grad } \psi_2 \times n) . n =$$

$$\text{curl } \frac{1}{\sigma d} \frac{\partial}{\partial t} (\text{grad } T \times n) . n \qquad (3)$$

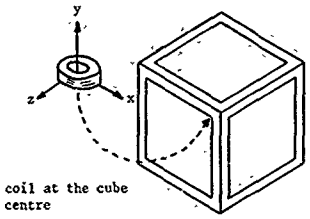
From equation (1) and since curl E = $-\frac{\partial B}{\partial t}$, then:

$$\text{curl } \frac{1}{\sigma d} (\text{grad } T \times n) . n = -B . n = \mu_1 \frac{\partial \psi_1}{\partial n} = \mu_2 \frac{\partial \psi_2}{\partial n} \qquad (4)$$

In order to implement this method, equation (3) must be solved on a conducting sheet, together with a Laplacian type equation in the surrounding non conducting volume.

The right hand side of equ (4) appears in a Galerkin treatment of div μ grad ψ, so this term links ψ_1 and ψ_2 to T, while yielding continuity of B.n across the sheet.

Figure 2 shows the experimental apparatus. Six copper sheets (280 x 280 x 0.55 mm) are symetrically arranged on the faces of a cubic (300 x 300 x 300 mm) wooden box. The copper sheets (conductivity 35.5 x 10^6 S/m) were not joined at the edges for this test as the resistivity of the joints would be difficult to quantify. Inside the box at the centre is a solenoidal coil (inner, outer radii 20 and 40 mm, thickness 15 mm), made from stranded wire and carrying 100 A turns.



line 1 from (0.03, 0.125, 0.185) to (0.21, 0.125, 0.185)

line 2 from (0.115, 0.1575, 0 03) to (0.115, 0.1575, 0.21)

line 3 from (0.115, 0.03, 0.185) to (0.115, 0.21, 0.185)

figure 2 The experimental apparatus

Results

Measurements were taken along the three lines shown on Figure 2. Results for 5, 10, 31 and 40 kHz are shown on Figures 3 - 5. The calculated values of B did not vary as much as the experimental values with frequency, so that only one representative calculated curve is shown on the figures. The calculated results tend towards the 5 kHz measured results, this , reasonable as the assumption of the 0.55 mm sheet being thinner than the skin depth is more valid at 5 kHz (0 84 mm) than at 40 kHz (0.3 mm). The correlation between measured and calculated values is quite good.

Conclusions

It has been shown that 3D finite elements should be useful in low frequency EMC calculations.

References

[1] D. Rodger and N. Atkinson 'Finite element method for 3D eddy current flow in thin conducting sheets' Proc IEE, Vol. 135, Pt A, No. 6, July 1988 pp 369 - 374.

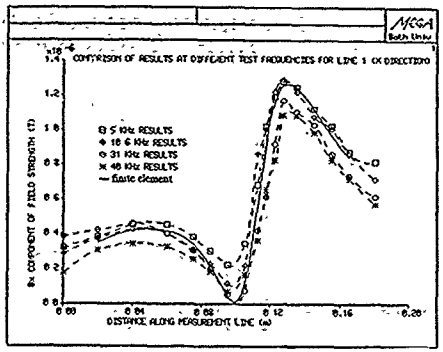


figure 3 Magnitude of B_x along line 1

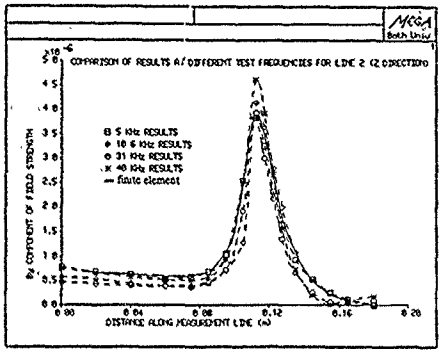


figure 4 Magnitude of B_y along line 2

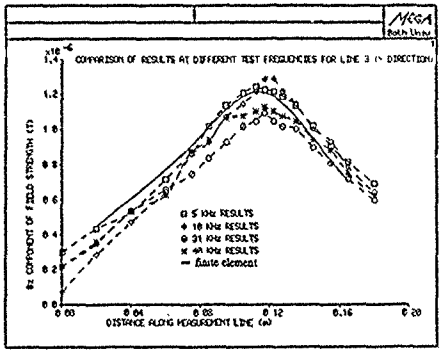


figure 5 Magnitude of B_z along line 3

3-D EDDY CURRENT ANALYSIS USING T-METHOD AND ITS APPLICATION TO NON-DESTRUCTIVE TESTING

T. TAKAGI, J. TANI
The Institute of Fluid Science
Tohoku University
Katahira 2-1-1, Sendai 980, Japan

K. MIYA
Nuclear Engineering Research Laboratory
The University of Tokyo
Tokai, Ibaraki 319-11, Japan

K. FUJIWARA, T. SAKAMOTO
System Engineering Division
Sumitomo Metal Industries, Ltd.
Nishinagasu 1-3, Amagasaki 660, Japan

Abstract-This paper describes numerical analysis using T-method and an application to eddy current testing. The code using twenty-node isoparametric elements was applied to a standard problem proposed by the TEAM Workshops and the results were compared with measured ones. The formulation using twelve-edge elements was also introduced and a new code has been developed.

I. INTRODUCTION

We have already proposed the T-method [1], where the magnetic scalar potential Ω of T-Ω method is not included. Advantages of the T-method are: (1) no variables in space and (2) easy treatment of external current and field. But it has a disadvantage that a large core memory is needed due to a dense matrix. In order to overcome this difficulty, we have already proposed an iterative solution technique for the T-method [1]. The code has been developed using twenty-node isoparametric elements and it was applied to an eddy current testing problem (a standard problem proposed by the TEAM Workshops [2]) for two frequencies (500Hz, 1kHz) [1].

In this paper we show improved results for the same eddy current problem and also present a new formulation using twelve-edge elements.

II. EDDY CURRENT ANALYSIS

Since current conservation is secured, current vector potential, T, is defined as, $J=\nabla\times T$. The relation between induced magnetic induction and current vector potential is given like [1],

$$B_e=\mu_0 T+\frac{\mu_0}{4\pi}\int_s T_n \nabla'\frac{1}{R}dS' \qquad (1)$$

where B_e and μ_0 are induced magnetic induction vector and magnetic permeability.

We obtain the governing equation for an AC eddy current problem using imaginary unit, j, and angular frequency, ω.

$$\nabla\times\frac{1}{\sigma}\nabla\times T+j\omega\mu_0 T+j\omega\frac{\mu_0}{4\pi}\int_s T_n\nabla\frac{1}{R}dS'+j\omega B_e=0 \qquad (2)$$

We must solve eq (2) with the Coulomb gauge (eq.(3)) and boundary condition (eq (4)) on conductor surface.

$$\nabla\times T=0 \qquad (3)$$
$$T\times n=0 \qquad \text{on conductor surface} \qquad (4)$$

In the previous paper [1] we solved eq.(2) and satisfied the Coulomb gage and the boundary condition by using a penalty method. When we used a large penalty number, good convergence characteristics could not be obtained in the iterative matrix solution.

Hence we here propose a new formulation using twelve-edge elements in order to avoid such difficulty. The current vector potential, T, is expressed in the local coordinates (ξ, η, ζ) using the shape functions proposed by K. Sakayama [3],

$$T=\{T_\xi, T_\eta, T_\zeta\}^T=\sum_{i=1}^{12} N_i T_i \qquad (5)$$

where $N_i=\{N_{\xi i}, N_{\eta i}, N_{\zeta i}\}^T$ These shape functions are defined like,

$$N_{\xi 1}=\frac{1}{4}(1-\eta)(1-\zeta),\quad N_{\xi 2}=\frac{1}{4}(1+\eta)(1-\zeta),$$
$$N_{\xi 3}=\frac{1}{4}(1+\eta)(1+\zeta),\quad N_{\xi 4}=\frac{1}{4}(1-\eta)(1+\zeta), \qquad (6)$$
$$N_{\xi 5}=N_{\xi 6}=N_{\xi 7}=N_{\xi 8}=N_{\xi 9}=N_{\xi 10}=N_{\xi 11}=N_{\xi 12}=0.$$

$N_{\eta i}$ and $N_{\zeta i}$ are expressed in the same manner. Using the above shape functions the Coulomb gage (eq.(3)) is automatically satisfied. The boundary condition (eq (4)) is also easily satisfied by prescribing the current vector potential on the surface to zero. The code has been developed here and it was applied to a simple 3-D problem. Fig.1 shows eddy current distribution (z=0.0, -0.0169m) in conductor when transient external field was applied. The results was favorably compared with those by the code "ELMES" using twenty-node elements.

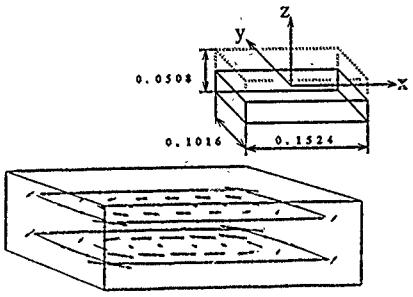


Fig 1 Eddy current density

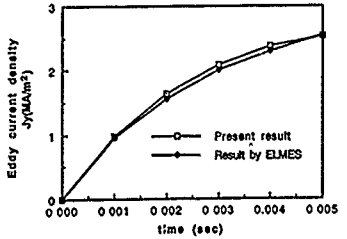


Fig 2 Time variation of eddy current density (x=0.073 , y=0.0 , z=0.0)

III. RESULTS AND DISCUSSION

ELMES was applied to an ECT problem. We adopted a standard one proposed in the TEAM Workshops [2]. The test piece is a rectangular block 330×285×30mm with a 40×0.5×10mm crack on the center of one of the larger faces. It is made of austenitic steel. Relative permeability μ_r is 1 and electrical conductivity σ is 0.14×10^6S/m.

A differential probe moves on the surface of the block. The probe is a cylinder with an active solenoid and two smaller receptive solenoids. The experiments were carried out for a parallel movement to the crack under four different frequencies (500Hz, 1kHz, 2kHz, 5kHz) [4].

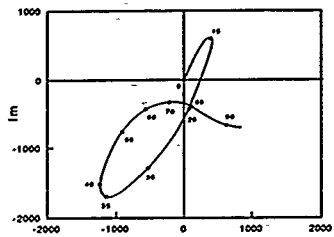
Numerical analysis was also carried out for the same problem. For the symmetry the region to be meshed is only a half of the block. The crack is treated as a low conductive region and included in the analysis domain. According to each coil position, we used a different type of mesh division. In order to get high accuracy of magnetic field calculation we adapted the mesh division to the positions of coils and a crack, and used the symmetric mesh division in the region near the coil.

Figs.3-4 show the numerical and experimental results for two frequencies. Table 1 shows the synthesis of the results. The notation in the table is same as that of Verite's [2] as shown in Fig.5. Numerical results of the University of Tokyo in the table were obtained from the references [1,5]. The results show almost good agreement with experimental ones both in shape (especially the location of turning points of the signal) and phase angle in the results of 500Hz and 1kHz.

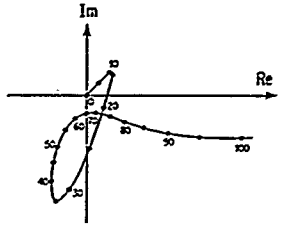
Authors would like to acknowledge Mr. S. Kawamura and Mr. T. Serizawa for their help in calculation. All calculation was performed using Cray-YMP8/4128 of the Supercomputer Center of the Institute of Fluid Science, Tohoku University.

REFERENCES

[1] T. Takagi et al., IEEE Trans. Mag. Vol.26(1990) pp.474-477.
[2] J. C. Verite, COMPEL, Vol.9(1990) pp.155-167.
[3] K. Sakiyama et al., IEEE Trans. Mag. Vol.26(1990) pp.1759-1761.
[4] T. Takagi et al., Proc. Asian TEAM Workshop and Seminar on Comput. Appl. Electromagn., IFS-TM001, Tohoku University (1991) pp.21-27.
[5] T. Takagi et al., COMPEL, Vol.9-A(1990) pp.260-262.

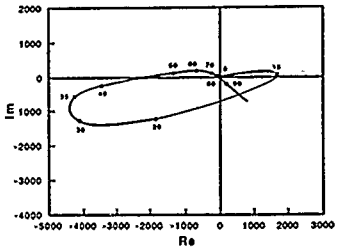


(a) Numerical result

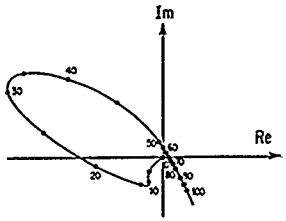


(b) Experimental result

Fig.3 Signal trajectories (500Hz)



(a) Numerical result



(b) Experimental result

Fig.4 Signal trajectories (5kHz)

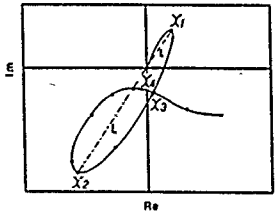


Fig.5 Definition of parameters

Table 1 The results of "Parallel movement to the crack"

Frequency	X1	X2	X3	X4	L8
f=500[Hz]					
Experiment (Univ. Tokyo)	10-15	35	20-25	70-75	3.2
Numerical Results (Tohoku Univ.)	15	35	15-20	70-80	2.9
Numerical Results (Univ. Tokyo)	10	35-40	30	75	1.9
f=1000[Hz]					
Exp.	15	35	15-20	85-90	3.5
Num.(Tohoku)	15	35	15-20	80-90	2.6
Num.(Tokyo)	10-20	35-40	20-30	80-90	1.7
f=2000[Hz]					
Exp.	10	30-35	NO	NO	5.5
Num.(Tohoku)	15	30-35	15-20	90-	2.6
f=5000[Hz]					
Exp.	10	30-35	NO	NO	5.9
Num.(Tohoku)	15	30-35	15-20	90-	2.6

FAST EVALUATION OF 3-D EDDY CURRENT LOSS DISTRIBUTION IN TRANSFORMER TANKS

J. TUROWSKI
SENIOR MEMBER IEEE
Institute of Electrical Machines and Transformers
Technical University of Lodz
ul. Stefanowskiego 18/22, 90-924 Lodz, Poland

Abstract – The author's model and method ROM-3D used so far for 3-D analysis of leakage field distribution has been now developed for a fast simplified evaluation of eddy-current loss distribution on internal surface of tank wall of large three-phase power transformers. Field components H_{mx} and loss distribution are calculated automatically, with analytic preprocessor and graphic postprocessor. Owing to eligible simplifi- cations, only 15-20 s on a PC computer are sufficient for simulation of one structure configuration. An influence of tank screening and general formula for the total power loss in tank wall is presented.

I. INTRODUCTION

Eddy currents induced in inactive parts of large three-phase power transformers produce highly non-uniform distributed losses and local heating, which are important factors of total energy saving and power system reliability. This distribution has extremely three-dimensional character, especially when various screens and shunts are applied. The most popular approach to solution of such problem is to apply 3-D finite element method [1]. However full three-dimensional solution, taking into account nonlinear magnetic permeability, heating, etc., has proved as "too costly in terms of both men-hours and computer time and not attractive for design use" [1]. Moreover, the 3-D FEM approach is not suitable for modelling on a personal computer, what is necessary in practice.

To overcome this problem the author's simplified reluctance network 3-D method [2,5] (Fig.2a), together with analytical formulae [4] was applied. Leakage field of three-phase power transformer in Fig.2 is simulated and resolved with the help of the ROM-3D computer program, containing dedicated analytical preprocessor, general network solver and graphic postprocessor. Such an approach enables to model and analyse field in one 3-D structural variant (configuration) within ca. 15 s CPU time on a personal computer IBM PC/AT. In the papers published so far [2,3] only one H_{mx} component was analysed. In the present paper both H_{mx} and H_{my} field components as well as resultant field $H_{ms} = \sqrt{H_{msx}^2 + H_{msy}^2}$ on the internal surface of tank wall were considered. The absolute value $|H_{ms}|^2$ is responsible for local losses.

II. LOCAL EDDY CURRENT LOSS IN STEEL WALL

Going out from Maxwell's equations for sinusoidal complex expressions and at μ = const

$$\nabla \times \bar{H}_m = \gamma \bar{E}_m \quad \text{and} \quad \nabla \times \bar{E}_m = j\omega\mu \bar{H}_m \qquad (1)$$

one can obtain [4] the classical formula for eddy current loss in a solid metal wall, W/m^2

$$P_1 = a_p \sqrt{\frac{\omega \mu_s}{2\gamma}} \cdot \frac{|H_{ms}|^2}{2} \qquad (2)$$

where at μ = const $a_p = 1$, but at steel with $\mu = \mu(H) \neq$ const $a_p \approx 1.4$ is the inward (in the Z direction) linearisation coefficient [4]. $\omega = 2\pi f$, $\mu_s = \mu_o \mu_{rs}(H_{ms})$ – magnetic permeability on steel surface; γ – metal conductivity.

To consider iron nonlinearity along surface one can use author's analytical approximation [4] of magnetisation curve (Fig.1) of solid steel

$$\sqrt{\mu_r}\, H^2 = c_1 H + c_2 H^2, \quad \text{for } 0 \le H \le 180 \cdot 10^2 \frac{A}{m} \qquad (3)$$

where $c_1 = 310 \cdot 10^2 \frac{A}{m}$ and $c_2 = 7.9$.

One can see from Fig.1 that even more simple approximation

$$\sqrt{\mu_r}\, H \approx 1.3\, c_2 H^2 = 10.27\, H^2 \qquad (4)$$

with corresponding error is possible.

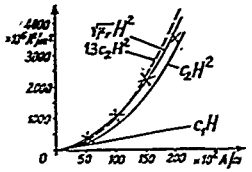


Fig.1. Analytical approximations (3) and (4) of magnetisation curve of solid steel. × × × – measured points.

III. COMPUTATION OF 3-D FIELD AND LOSS DISTRIBUTION ON TANK WALL

Leakage field in a three-phase transformer has been simulated with the 3-D reluctance network (Fig.2a).

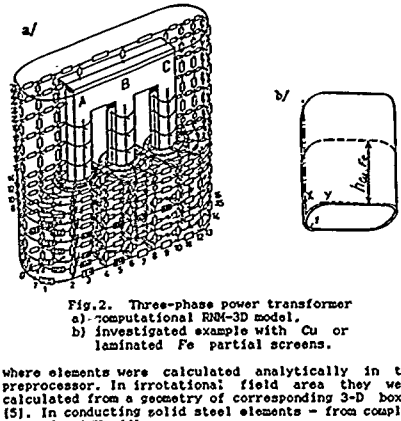


Fig.2. Three-phase power transformer a) computational ROM-3D model, b) investigated example with Cu or laminated Fe partial screens.

where elements were calculated analytically in the preprocessor. In irrotational field area they were calculated from a geometry of corresponding 3-D boxes [5]. In conducting solid steel elements – from complex value in 1/H [4]

$$B_{\mu 1} \approx (0.37 + j\,0.61) \sqrt{\frac{\omega \gamma}{\mu_s}} \qquad (5)$$

which takes into account nonlinear steel permeability and eddy current reaction. Copper or aluminium screens were also represented by corresponding complex formulae [3]. The model in Fig.2a has 4 × 186 nodes. Some more in windows than that in papers [3] and [6].

As examples, tanks without or with continuous Cu or laminated Fe screens (Fig.2b) were considered. Edge of the screen is situated in discrete node points. The reluctance on the edge is calculated as parallel connection of two components

$$R'_{\mu Cu} = 2R_{\mu Cu} \gg R_{\mu 1} \quad \text{and} \quad R'_{\mu 1} = 2R_{\mu 1}$$

Therefore the first of these values was neglected, and the second - a little decreased as $R'_{\mu 1} \approx R_{\mu 1}$, giving a relatively small error. One rectangular element was subdivided into four small rectangles. Field and loss in each of them were calculated from the adjacent components H^2_{mx} and H^2_{my}. In this way the network was practically made four time denser than in previous works [3,6].

A total loss computation should consider all kinds of elements, according to (2), and corresponding screening coefficients p_e and p_m for electromagnetic and magnetic screens respectively. The total losses in W are therefore

$$P = \frac{1}{2}\sqrt{\frac{\omega \mu_o}{2\gamma_{Fe}}} \cdot \left[p_e \iint_{A_e} \sqrt{\mu_{rse}}\, H^2_{ms}\, dA_e + \right.$$

$$\left. + p_m \iint_{A_m} \sqrt{\mu_{rsm}}\, H^2_{ms}\, dA_m + e_p \iint_{A_{St}} \sqrt{\mu_{rs}}\, H^2_{ms}\, dA_{St} \right] \quad (6)$$

where it was supposed

$$\sqrt{\mu_{rse}} \approx \sqrt{\mu_{rsm}} \approx \sqrt{\mu_{rs}}$$

$$p_e \approx \frac{1}{\gamma_{Cu}\, d}\sqrt{\frac{\gamma_{Fe}}{\omega \mu_o \mu_{rs}}} \approx \frac{0.2 \cdot 10^{-3}\ \text{m}}{d_{Cu}}$$

$$p_m \approx \frac{18 \cdot 1.4}{(4.2\pi n)^2 \sqrt{n^2}}; \qquad e_p \approx 1.4 \quad (7)$$

d_{Cu} - thickness of Cu screen in m, n - number of sheets in flat laminated magnetic screen, eg. n = 15.

Influence of screens on the field and losses is shown in Figures 3 to 5.

The loss distributions p^r in Figs. 3c, 4c, 5c are expressed in relative units referred to

$$\frac{1}{2}\sqrt{\frac{\omega \mu_o}{2\gamma_{Fe}}}\, H^2_{mp}, \quad \text{where} \quad H_{mp} = \frac{\sqrt{2}\, I^N_{NH}}{h_R} \quad (8)$$

I^N_{NH} - rated ampere-turns of HV winding; $h_R = h/K$, h - windings height, $K = 1 - (a_1+a_2+\delta)/(\pi h)$ - Rozowski's coefficient [5].

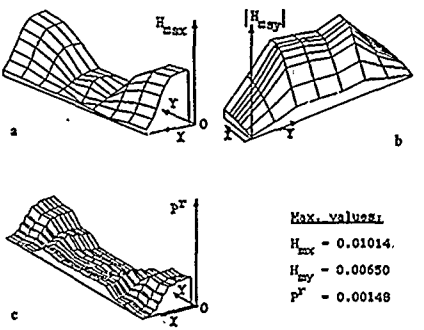


Fig.3. Field components H_{mx} (a), H_{my} (b), and eddy current losses (c) on tank surface of a 315 kVA transformer, without screens.

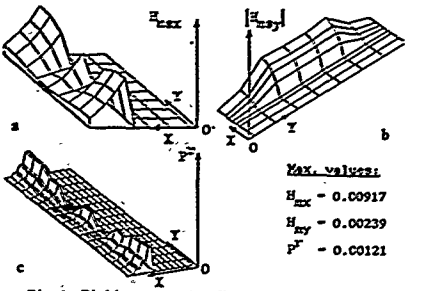


Fig.4. Field components H_{mx} (a), H_{my} (b), and eddy current losses (c) on tank surface of a 315 kVA transformer with the partial Fe screen.

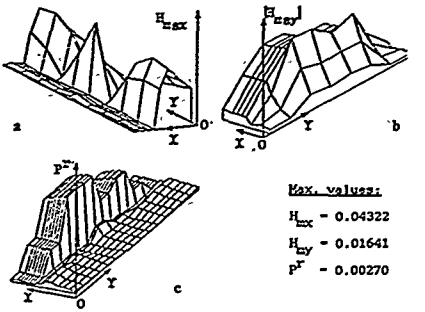


Fig.5. Field components H_{mx} (a), H_{my} (b), and eddy current losses (c) on tank surface of a 315 kVA transformer with the partial Cu screen.

IV. CONCLUSION

Application of the network approach with the RNM-3D computer program has shown high effectiveness of this method at evaluation of 3-D field and eddy current loss modelling and analysis. A lot of information can be obtained within seconds on a personal computer.

Acknowledgement. The author expresses his thanks to Marek Turowski, M.Sc. and Mirosław Kopeć, M.Sc., both Electronic Engineers, for their assistance in modification of the RNM-3D program and computations.

The work was made within the project EDT-3D sponsored by Polish Ministry of National Education.

V. REFERENCES

1. Coulson M.A., Preston T.W., Reece A.B.J.: "3-Dimensional Finite-Element Solvers for the Design of Electrical Equipment". *COMPUMAG 1985*. Fort Collins, USA.
2. Turowski J.: "Modelling and Simulation of Electromagnetic Field in Electrical Machines and Transformers with the Help of Equivalent Reluctance Network". *IMACS Annals*, Vol.6, Basel, 1989.
3. Turowski J., Kopeć M., Turowski M.: Method of Fast Analysis of 3-D Leakage Fields in Large Three-Phase Transformers". *3DMAG'89* and *COMPEL*, Vol.9-A, 1990.
4. Turowski J.: "Obliczenia elektromagnetyczne elementów maszyn i urządzeń elektrycznych". (book in Polish) WNT, Warszawa, 1982.
5. Turowski J.: "Model reluktancyjny pola rozproszenia transformatorów". *Rozprawy Elektrotechniczne*, vol.30, No.4, 1984, pp. 1121-1144.
6. Turowski J., Kopeć M., Turowski M.: "Fast 3-D Analysis of Combined Cu-Fe Screens in Three-Phase Transformer. *ISEF'89* and *COMPEL* 1990.

AN ITERATIVE METHOD FOR SOLVING THE INVERSE PROBLEM OF ELECTRICAL LOG

KANG ZHENG AND LIU JIA QI
Dept. of Math., Harbin Institute of Technology, China

Abstract: In this paper, the problem of electrical log in geophysics is reduced to the inverse problem of an elliptic differential equation. An iterative method for solving the inverse problem is presented.

I. INTRODUCTION

The technique of eletrical log is very important to production of petroleum in oil field. It is put into practice by means of a system of electrodes below well. The electrodes transmit steady currents to the earth so that an electrical field is created. By receiving the electric responses, one can identify various formation. If the earth is symmetrized with respect to the well axis, the problem of electric log can be reduced to a two-dimensiona problem in a meridian plane. The distribution of underwell midium is demonstrated in Fig.1, where $\Omega_0, \Omega_1, \Omega_2$ and Ω_3

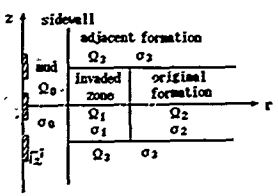


Fig.1

represent the respective area of muddy zone, invaded zone, original formation and adjacent formation. Let $\sigma(r,z)$ be distributive function of electric conductivity. It can be assumed to be a piecewise constant function [1], i.e.

$$\sigma(r,z)=\sigma^{(i)}, \quad (r,z)\in\Omega_i,\ i=0,1,2,3.$$

where $\sigma^{(0)}, \sigma^{(1)}, \sigma^{(2)}$ and $\sigma^{(3)}$ are the respective conductivity of mud, invaded zone, original formation and adjencent formation. $\sigma^{(0)}$ is known, $\sigma^{(1)}, \sigma^{(2)}$ and $\sigma^{(3)}$ are unknown parameters characterizing the unknown formation in the electric field. Generally, the electrode's currents and the electric potential on the electrode's surface can be measured. The inversion of electric log is to recover $\sigma^{(1)}, \sigma^{(2)}$ and $\sigma^{(3)}$ by electrode's and potential values.

II. MATHEMATICAL MODEL

The potential distribution function $u(r,z)$ in the underwell electric field satisfies the following partial differential equation [1]

$$\frac{\partial}{\partial r}(K\frac{\partial u}{\partial r})+\frac{\partial}{\partial z}(K\frac{\partial u}{\partial z})=0, \qquad (r,z)\in\Omega=\bigcup_{i=0}^{3}\Omega_i \quad (1)$$

and boundary conditions

$$\frac{\partial u}{\partial n}=0, \quad (r,z)\in\Gamma_1 \qquad (2)$$

$$u=b_i, \ (r,z)\in\Gamma_2^i, \ i=1,\cdots,p \qquad (3)$$

$$u=0. \quad (r,z)\in\Gamma_3 \qquad (4)$$

where $K=r\sigma$; Γ_1 is the insulation boundary, n is the outer normal direction of Γ_1; Γ_2^i is the ith electrode's boundary, b_i is the potential on Γ_2^i, p is the number of the electrodes; Γ_3 is the infinite boundary. Notate $\Gamma_2=\bigcup_{i=1}^{p}\Gamma_2^i$ then Γ_1, Γ_2 and Γ_3 form the whole boundary of Ω.

Let Γ_c be the demarcation plane between the zones of respective medium. On Γ_c, there exist join conditions

$$u_+=u_-, \quad (\sigma\frac{\partial u}{\partial n})_+=(\sigma\frac{\partial u}{\partial n})_-, \quad (r,z)\in\Gamma_c \qquad (5)$$

where n is Γ_c's normal direction and subscript '+', '−' represent two sides of Γ_c.

In addition, u also satisfies measured conditions (current conditions)

$$\int_{\Gamma_2^i}\sigma\frac{\partial u}{\partial n}ds=I_i, \quad i=1,\cdots p \qquad (6)$$

where $I_i (i=1,\cdots p)$ are the known current values.

Thus, the model of electrical log is reduced to the inverse problem constituted by eq.(1), boundary conditions (2)~(4), join conditions (5) and measurement condition (6).

III. ITERATIVE METHOD

In order to determine σ, the iterative procedure is introduced by letting [2]

$$k_{n+1}=k_n+\delta k_n, \ u_{n+1}=u_n+\delta u_n, \ \delta k_n=r\delta\sigma_n, \quad n=0,1\cdots \qquad (7)$$

where σ_0 is the initial guess. $\|k_n\|\gg\|\delta k_n\|$ and $\|u_n\|\gg\|\delta u_n\|$ are necessary for the convergence of the iterative algorithm. Upon substituting (7) into (1), one obtains

$$\frac{\partial}{\partial r}((k_n+\delta k_n)\frac{\partial(u_n+\delta u_n)}{\partial r})+\frac{\partial}{\partial z}((k_n+\delta k_n)\frac{\partial(u_n+\delta u_n)}{\partial z})=0 \qquad (8)$$

Upon collecting equal order terms and neglecting second order terms of δ, one obtains a system for u_n

$$\begin{cases}\dfrac{\partial}{\partial r}(k_n\dfrac{\partial u_n}{\partial r})+\dfrac{\partial}{\partial z}(k_n\dfrac{\partial u_n}{\partial z})=0\\ \dfrac{\partial u_n}{\partial n}=0, \quad (r,z)\in\Gamma_1\\ u_n=b_i, \ (r,z)\in\Gamma_2^i,\\ u_n=0, \quad (r,z)\in\Gamma_3\end{cases} \qquad (9)$$

and a system for δu_n

$$\frac{\partial}{\partial r}(k_n\frac{\partial\delta u_n}{\partial r})+\frac{\partial}{\partial z}(k_n\frac{\partial\delta u_n}{\partial z})=$$

$$\begin{cases} -\left(\frac{\partial}{\partial r}(\delta k_\sigma \frac{\partial u_\sigma}{\partial r}) + \frac{\partial}{\partial z}(\delta k_\sigma \frac{\partial u_\sigma}{\partial z})\right) \\ \frac{\partial \delta u_\sigma}{\partial n} = 0, \quad (r,z) \in \Gamma_1 \\ \delta u_\sigma = 0, \quad (r,z) \in \Gamma_2 \cup \Gamma_3 \end{cases} \tag{10}$$

By using the Green's function method, (10) can be transform into and explicit fredholm integral equation of the first kind for the unknown δk_σ

$$\delta u_\sigma(r,z) = \int_\Omega \int G_\sigma(r,z,r',z') W_\sigma(r',z') dr'dz' \tag{11}$$

where $W_\sigma(r,z) = -[\frac{\partial}{\partial r}(\delta k_\sigma \frac{\partial u_\sigma}{\partial r}) + \frac{\partial}{\partial z}(\delta k_\sigma \frac{\partial u_\sigma}{\partial z})]$, and $G_\sigma(r,z,r',z')$ satisfies.

$$\begin{cases} [\frac{\partial}{\partial r}(k_\sigma \frac{\partial}{\partial z}) + \frac{\partial}{\partial r}(k_\sigma \frac{\partial}{\partial z})] G_\sigma(r,z,r',z') = \delta(r-r', z-z') \\ \frac{\partial G_\sigma}{\partial n} = 0, \quad (r,z) \in \Gamma_1 \\ G_\sigma = 0, \quad (r,z) \in \Gamma_2 \cup \Gamma_3 \end{cases} \tag{12}$$

Approximately replacing δu_σ in (11) by $u-u_\sigma$ and taking the normal derivative with respect to Γ_2^1, one obtains

$$\frac{\partial u}{\partial n} - \frac{\partial u_\sigma}{\partial n} = \int_\Omega \int \frac{\partial G_\sigma(r, z, r', z')}{\partial n} W_\sigma(r',z') dr'dz', \quad (r,z) \in \Gamma_2^1, \tag{13}$$

By integrating (13) on Γ_2^1 and according to (6),(13) is written as

$$I_1 - \int_{\Gamma_2^1} \frac{\partial u_\sigma}{\partial n} ds = \int_\Omega \int Q_1^1(r',z') W_\sigma(r',z') dr'dz', \tag{14}$$

where $I_1 = I_1 / \sigma^{(0)}$, and

$$Q_1^1(r',z') = \int_{\Gamma_2^1} \frac{\partial G_\sigma(r,z,r',z')}{\partial n} ds \tag{15}$$

The right hand side of (14) can be further simplified

$$\begin{aligned} &\int_\Omega \int Q_1^1(r',z') W_\sigma(r',z') dr'dz' \\ &= \sum_{j=0}^{3} \int_{\Omega_j} \int Q_1^1(r',z') W_\sigma(r',z') dr'dz' \\ &= -\sum_{j=0}^{3} \delta \sigma_\sigma^{(j)} \int_{\Omega_j} \int Q_1^1(r',z') [\frac{\partial}{\partial r}(r \frac{\partial u_\sigma}{\partial r}) + \frac{\partial}{\partial z}(r \frac{\partial u_\sigma}{\partial z})] dr dz \\ &= -\sum_{j=1}^{3} C_{1\sigma}^j \delta \sigma_\sigma^{(j)}, \quad (\text{as } \delta \sigma_\sigma^{(0)} = 0) \end{aligned}$$

where $C_{1\sigma}^j = -\int_{\Omega_j} \int Q_1^1(r',z') [\frac{\partial}{\partial r}(r \frac{\partial u_\sigma}{\partial r}) + \frac{\partial}{\partial z}(r \frac{\partial u_\sigma}{\partial z})] dr dz$.

Let $d_\sigma^1 = I_1 - \int_{\Gamma_2^1} \frac{\partial u_\sigma}{\partial n} dz$, $(i=1,2,\cdots,p)$. Then, (14) can be further written as

$$-\sum_{j=1}^{3} C_{1\sigma}^j \delta \sigma_\sigma^{(j)} = d_\sigma^1, \quad i=1,\cdots, p \tag{16}$$

(16) is a system of linear algebraic equations with respect to $\delta \sigma_\sigma^{(j)}$. Solving (16), one can obtain $\delta \sigma_\sigma^{(j)}$.

Now, eqs.(7),(9) and (16) form the basic structure for each iteration in the numerical iterative algoritha.

In order to test the feasibility of the method, the following numerical simulation procedure is carried out:

First, one chooses parameter $\sigma(r,z)$ which supposed to represent the correct conductivity of an earth and also the boundary conditions b_i, $i=1,\cdots, p$. Then the system (1)~(5) with the chosen $\sigma(r,z)$ and b_i, $i=1,\cdots, p$ is solved by numerical method. thus one can generate the measured data I_i, $i=1,\cdots, p$. Next, the initial guess $\sigma_\sigma(r,z)$ is assumed. Upon solving (7),(9) and (16), σ_1 is obtained. Then in a similar manner σ_2 is obtained. One continues this procedure until finally a numerical limit σ_s is reached. The compctational results for various cases are tabulated in the following table

ex.	n	initial value			truth value			iterative solution		
		$\sigma_\sigma^{(1)}$	$\sigma_\sigma^{(2)}$	$\sigma_\sigma^{(3)}$	$\sigma^{(1)}$	$\sigma^{(2)}$	$\sigma^{(3)}$	$\sigma_s^{(1)}$	$\sigma_s^{(2)}$	$\sigma_s^{(3)}$
1	5	2	2	2	10	15	20	10.000	15.00	19.98
2	8	5	5	5	50	100	10	50.000	99.82	10.00
3	4	1	1	1	3	10	1	3.00	9.99	1.00
4	4	1	-1	1	5	20	3	5.09	19.80	3.00

The simulations show that the iterative method for solving the inverse problem of electrical log is effective.

REFERENCES

[1] D.Q.Li,Apply the finite element method to electrical log,oil industry publishing house. 1980 (in Chinese)

[2] Y.M.Chen,J.comput.phys,vol.43,No.2.1981

CONVEX-ANALYSIS AND SUBDIFFERENTIAL METHODS IN ELECTROMAGNETISM

Stanisław K. Krzemiński
Institute Electrical Theory and Measurements
Warsaw Technical University
Koszykowa 75, 00-661 Warsaw

Abstract: A new method of formulating boundary and initial boundary value problems with unilateral constraints for field vectors is presented. Such problems of electrodynamics were described with a system of multiequations and variational inequalities.

I. INTRODUCTION

In the last years much was done in application of mathematical physics equations to field problems as regards mathematical modelling of initial boundary value problems with constraints for field vectors. The functions of state satisfy not only the conditions resulting from the given boundary value problem but also additional ones entailed by unilateral constraints, that is inequality conditions.

Such problems may be described naturally with elements of convex analysis and subdifferential calculus [1,3]. This approach to problems with unilateral constraints results in their being described with multiequations, or differential inclusions or variational inequalities [1,2]. This paper presents a mathematical model for an initial boundary value problem with unilateral constraints for the source current density vector and the magnetic induction vector.

ELEMENTS OF SUBDIFFERENTIAL AND CONVEX ANALYSIS

The mathematical basis of convex analysis and subdifferential calculus are given in [3]. Here will be given only those elements of this theory which are relevant to the subsequent considerations.

(D1): If functional $J:V\longrightarrow R$ is convex and lower semicontinuous, then its coupled functional, or, in other words, the polar functional $J^*:V^*(\Omega)\longrightarrow R$ is of the following form,

$$J^*(u^*) = \sup_{u\in V}\{\langle u^*,u\rangle - J(u)\} \qquad (1)$$

(D2): We call the element $u^*\in V^*(\Omega)$ a subgradient of the convex lower semicontinuous functional $J:V\longrightarrow R$ in point (u) when it satisfies the Fenchel inequality [3].

$$\langle u^*, v-u\rangle + J(u) \le J(v) \quad ,\forall v\in V \qquad (2)$$

We denote the set of subgradients (u^*) of functional $J:V\longrightarrow R$ by $\partial J(u)$. The subdifferential $\partial J(u)$ has the following properties.

$$u^*\in \partial J(u) \leftrightarrow u\in \partial J^*(u^*) \qquad (3)$$

$$\partial(J\circ L)(u) = L^*\partial J(L\circ u) \qquad (4)$$

An initial boundary value problem with unilateral constraints for field vectors will now be formulated.

MODEL INITIAL BOUNDARY VALUE PROBLEM OF ELECTRODYNAMICS

A model of initial boundary valueproblem for a medium with dominant current conductance properties will be considered. For such medium the canonical equations [2] are given by

$$L^*\circ H = j + \gamma\frac{\partial A}{\partial t} \qquad (5)$$

$$B = \mu\circ H \qquad (6)$$

$$L\circ A = B \qquad (7)$$

where $L^*=L=(\nabla\times)$ are the operators. The vector of source current density $j(x,t)$ and the vector of magnetic field intensity $H(x,t)$ have two components,

$$j(x,t) = \hat{j}(x,t) + \bar{j}(x,t) \qquad (8)$$

$$H(x,t) = \hat{H}(x,t) + \bar{H}(x,t) \qquad (9)$$

Components $\hat{j}$ and $\hat{H}$ are nonlinear and depend on the vector B. Constraints for the induction vector are contained in the defined set $K(\Omega)$,

$$K(\Omega) = \{ A: A\in V(\Omega), |\nabla\times A| \le B_o \} \qquad (10)$$

The indicator function of set $K(\Omega)$ is of the form,

$$\mathrm{ind}_K(A) = p(A) = \begin{cases} 0 & \text{for } A\in K(\Omega) \\ +\infty & \text{for } A\notin K(\Omega) \end{cases} \qquad (11)$$

The normal cone of set $K(\Omega)$ in point A is a set of subgradients A^* or $\bar{j}$ of the form. Accordingly, one may write that the controlled component of source current $\bar{j}$ belongs to the cone.

$$\bar{j}\in -N_K(A) = -\partial\mathrm{ind}_K(A) = -\partial p(A) \qquad (12)$$

The characteristic of the magnetic medium, $H(B)$ is given by,

$$H(B) = \begin{cases} -\nu_o B & \text{for } B<0 \\ \nu_o B & \text{for } 0\le B\le B_o \\ \nu_o B^2 & \text{for } B>B_o,\ \nu_o = B_o/H_o \end{cases} \qquad (13)$$

$$\mu^{-1} = \frac{\partial H}{\partial B} \in \partial H(B) \qquad (14)$$

The nonlinear component $\bar{H}(B)$ belongs to the subdifferential of the indicator function of the epiH.

$$\bar{H}(B) \in \partial ind_{epiH}\{B,H(B)\} = \partial \varphi(B) \qquad (15)$$

Thus,the dynamics of electromagnetic field in region Ω with unilateral constraints for field vectors contained in (12) and (15) was described by two inclusions, multiequations of the following form.

$$L^* \cdot H - j + \gamma \frac{\partial \lambda}{\partial t} \in -\partial \varphi(A) \qquad (16)$$

$$\mu^{-1} \cdot L\,A = H \in \partial \varphi(L\,A) \qquad (17)$$

The subdifferential $\partial H(B)$ of the characteristic (13) has the following analytical form,

$$\partial H(B) = \begin{cases} -2\nu_o & \text{for } B < 0 \\ [-2\nu_o, \nu_o] & \text{for } B = 0 \\ \nu_o & \text{for } 0 < B < B_o \\ [\nu_o, 2\nu_o] & \text{for } B = B_o \\ 2\nu_o B & \text{for } B > B_o \end{cases} \qquad (18)$$

Fig.1. presents the subdifferential of the characteristic H(B).

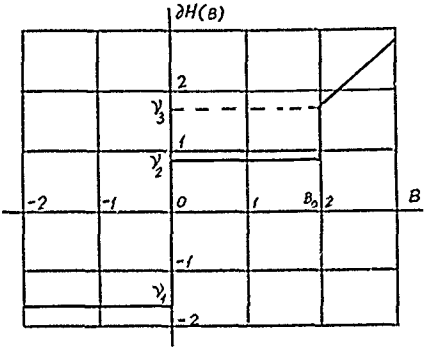


Fig.1. Subdifferential $\partial H(B)$

In virtuo of the subdifferential's property (D2) the system of multiequations (16)-(17) is equivalent to a system of variational inequalities (19),(20) [1], in which (φ) and (ψ) are indicator functions of sets $K(\Omega)$ and epiH respectively.

$$\langle U-A, L^* \cdot H - j + \gamma \frac{\partial \lambda}{\partial t} \rangle + \varphi(U) - \varphi(A) \le 0 \qquad (19)$$

$$\langle V - H\,,\ L \cdot A \rangle + \psi(V) - \psi(H) \le 0 \qquad (20)$$

Eliminating the vector of magnetic field intensity H from (20), one gets the parabolic type evolution variational inequality,

$$\langle \gamma \frac{\partial \lambda}{\partial t}\ ,U-A \rangle + a(A,U-A) \le \langle j,U-A \rangle \qquad (21)$$

satisfied for every vector $U \in K(\Omega)$ and $A \in K(\Omega)$.

INEQUALITY MODEL OF THE BOUNDARY VALUE PROBLEM

The model initial boundary value problem of electrodynamics is of the form,for given current sources j, τ and material parameters γ, μ in region Ω, determine the potential distribution $A \in K(\Omega)$ satisfy- ing variational inequality (22) and the initial condition (23),

$$\langle \gamma \frac{\partial A}{\partial t}\ ,U-A \rangle + a(A,U-A) \le \langle j,U-A \rangle \qquad (22)$$

$$A(x,0) = A_o(x),\ x \in \Omega, \forall U \in K(\Omega) \qquad (23)$$

In the case of a magnetostatic field, the boundary value problem is, for given current sources j, τ and material parameters γ, μ in region (Ω) determine distribution of magnetic potential $A \in K(\Omega)$ satisfying the variational inequity

$$(A,U-A) \le \langle j,U-A \rangle\ ,\quad \forall\ U \in K(\Omega) \qquad (24)$$

The variational inequality (24) was solved by approximating function A(x,y), $(x,y) \in \Omega$, with a finite element.

CONCLUSIONS

Subdifferential calculus and convex analysis have never previously been applied to descriptions of nonlinear problems of electrodynamics with unilateral constraints for field vectors.

This approach enables, in a natural manner, the construction of a mathematical model consisting of elliptic and parabolic variational inequalities.

REFERENCES

[1] G. Duvaut, J. L Lions," Inequalities in Mechanics and Physics, "Springer -Verlag, Berlin, 1976.

[2] J.T.Oden, J.N.Redy,"On Dual Complementary Variational Principles in Mathematical Physics, Inter.Jour.of Eng.Science, vol.12 pp.1-29, 1974.

[3] R.T. Rockafellar, "Convex Analysis," Princeton University Press, 1970.

FIXED POINT THEOREMS AND N-BODY PROBLEM OF CLASSICAL ELECTRODYNAMICS

Vasil G. Angelov

Mining and Geological University
Department of Mathematics
1156 Sofia, Bulgaria

Abstract. N-body problem of classical electrodynamics is considered. The Newtonian instantaneous action at a distance is replaced by an action at a distance propagated with finite velocity. Fixed point approach allows to formulate "escape" conditions for the charged particles.

I. INTRODUCTION

The two-body problem of classical electrodynamics without radiation term has been posed by J.L.Synge [1], using the relativistic expressions for the Lorentz ponderomotive force, due to W.Pauli [2]. The main feature of his considerations is that the Newtonian instantaneous action at a distance is replaced by an action at a distance propagated with finite velocity. The problem stated has been investigated by several authors using different approaches (cf. [3]-[5]).

A natural extension of the problem mentioned is a N-body problem of classical electrodynamics. From the physical point of view this problem is based on the linear superposition principle of the fields. This means that the Lorentz force can be calculated by summation of the retarded fields produced by all other particles. In [6] one-dimensional case it has been considered.

In the present paper we formulate an initial value problem for equations of motion corresponding to three-dimensional N-body problem. By means of a fixed point theorem, proved in [5], we formulate "escape" conditions for the moving charged particles.

II. FIXED POINT THEOREMS

Let X be a separated uniform space with a uniformity generated by a saturated family of pseudometrics $\mathcal{A} = \{\rho_\beta(x,y): \beta\in\mathcal{B}\}$ where $\mathcal{B}$ is an index set. Let $\varphi:\mathcal{B}\to\mathcal{B}$ be a mapping with iterates defined as follows: $\varphi^k(\beta) = \varphi(\varphi^{k-1}(\beta)), \varphi^0(\beta)=\beta$ $(k=1,2,\dots)$. Let (Φ) be a family of contractive functions $\Phi_\beta(t): R^1_+\to R^1_+$ with the properties:

$(\Phi 1)$ $\Phi_\beta(t)$ is continuous from the right, strictly increasing and $0<\Phi_\beta(t)<t$ for $t>0$;

$(\Phi 2)$ $\lim_{n\to\infty} \Phi_\beta(\Phi_{\varphi(\beta)}(\dots\Phi_{\varphi^n(\beta)}(t)\dots)) = 0$ for every $t>0$ and $\beta\in\mathcal{B}$.

The operator $T: X \to X$ is said to be Φ-contractive if for every $x,y\in X$ and $\beta\in\mathcal{B}$ the inequality

$$\rho_\beta(Tx,Ty)\le\Phi_\beta(\rho_{\varphi(\beta)}(x,y))$$

is satisfied.

THEOREM 2.1 [5] Let $T: X \to X$ be Φ-contractive. For each $\beta\in\mathcal{B}$ there is $\bar\Phi_\beta(t)\in(\Phi)$ such that for $t\ge0$ $\sup\{\Phi_{\varphi^k(\beta)}(t): k=0,1,2,\dots\}\le\bar\Phi_\beta(t)$ and $\bar\Phi_\beta(t)/t$ is non-decreasing. In addition, there is $x_0\in X$ such that

$$\rho_{\varphi^k(\beta)}(x_0,Tx_0)\le Q(\beta,x_0,Tx_0)<\infty$$

$(k=0,1,2,\dots)$.

Then T has at least one fixed point in X.

It is easy to verify that a sum of Φ-contractive and completely continuous operator is no longer densifying one, because of that we use the notion of Φ-densifying operator, introduced in [5].

The operator T is said to be Φ-densifying if for every bounded set $\Omega\subset X$ and $\beta\in\mathcal{B}$

$$\gamma_\beta(T(\Omega))\le\Phi_\beta(\gamma_{\varphi(\beta)}(\Omega))$$

where γ_β coincides with Kuratowskii's or Hausdorff's measure of noncompactness.

It is known that every separated locally convex (linear topological) space possesses a uniformizable topology being completely regular. We suppose E has the property (C): the convex closure of every compact set is compact in E. When E is a Banach space (C) is satisfied in view of Mazur's theorem. If E is a locally convex one, (C) is satisfied if E is complete or even quasycomplete.

THEOREM 2.2 [5] Let $T:M\to M$ be continuous Φ-densifying mapping. For every bounded set Ω with $\gamma_\beta(\Omega)>0$ the inequality holds

$$\gamma_{\varphi^n(\beta)}(\Omega)\le Q(\beta,\Omega)<\infty \quad (n=0,1,2,\dots)$$

for some positive constant Q. Then T has at least one fixed point in M.

III. ON THE N-BODY PROBLEM

We formulate N-body problem taking into account that every charged particle is under the influence of another N-1 particles. The right hand sides of the equations of motion we calculate by Lienard-Wiechert retarded potentials following the technique due to Synge [1] and Pauli [2]. So the right hand side of every equation turns out a sum of retarded fields produced by the last N-1 particles.

Introduce the denotations (c.f.[5]): the space-time coordinates are $\mathfrak{X}_\kappa=(x^\nu, x_4=ict)$ (c - the speed of light), where Latin suffices run over 1-4, while Greek suffices run over 1-3 with Einstein summation convention; the scalar product is $\langle\vec a,\vec b\rangle=a_\kappa b_\kappa$; the proper masses are $m_\kappa (l=1,2,\dots,N)$, the charges e_κ, unit tangent vectors $\lambda^{(\kappa)}_\mu$, the elements of the

proper time $ds^{(\kappa)}$, the velocities $u_\alpha^{(\kappa)}$; $u_\alpha^{(\kappa)} u_\alpha^{(\kappa)} = c^2$ ($\sum_\alpha (u_\alpha^{(\kappa)})^2$),

$$\lambda_\alpha^{(\kappa)} = \frac{u_\alpha^{(\kappa)}}{\sqrt{c^2 - u_\alpha^{(\kappa)} u_\alpha^{(\kappa)}}}, \quad \lambda_4^{(\kappa)} = \frac{ic}{\sqrt{c^2 - u_\alpha^{(\kappa)} u_\alpha^{(\kappa)}}},$$

$Q^{(\kappa,n)} = \frac{e_\kappa e_n}{m_\kappa}$, $\xi_r^{(\kappa,n)}$ are isotopic vectors

$$\xi_r^{(\kappa,n)} = \left\{ x_\alpha^{(\kappa)}(t) - x_\alpha^{(n)}(t - \tau_{\kappa n}(t)),\ ic\tau_{\kappa n}(t) \right\}$$

where the retardations $\tau_{\kappa n}(t)$ satisfy the functional equations

$$\tau_{\kappa n}(t) = \frac{1}{c} \sqrt{\sum_{\alpha=1}^{3} \left[x_\alpha^{(\kappa)}(t) - x_\alpha^{(n)}(t - \tau_{\kappa n}(t)) \right]^2}$$

Equations of motion for N charged particles are:

$$m_\kappa \frac{d\lambda_n^{(\kappa)}}{ds^{(\kappa)}} = \frac{e_\kappa}{c^2} \left[F_{n\ell}^{(\kappa,1)} + \dots + F_{n\ell}^{(\kappa,\kappa-1)} + F_{n\ell}^{(\kappa,\kappa+1)} + \dots + F_{n\ell}^{(\kappa,N)} \right] \lambda_\ell^{(\kappa)} \quad (1)$$

(k=1,2,...,N), where $F_{n\ell}^{(\kappa,m)} = \frac{\partial \Phi_\ell^{(m)}}{\partial x_n^{(\kappa)}} - \frac{\partial \Phi_n^{(m)}}{\partial x_\ell^{(\kappa)}}$

and $\Phi_\ell^{(m)} = -\frac{e_m \lambda_\ell^{(m)}}{\langle \lambda^{(m)}, \xi^{(\kappa,m)} \rangle}$ -retarded potentials.

The system (1) is a neutral functional differential one with respect to the unknown velocities. After usual substitutions we can reduce (1) to the following one:

$$\begin{aligned} w_\alpha^{(\kappa)}(t) &= \sum_{\substack{m=1 \\ m \neq \kappa}}^{N} F_\alpha^{(\kappa,m)}, \quad t \in [0,\infty) \\ w_\alpha^{(\kappa)}(t) &= w_{\alpha 0}^{(\kappa)}(t), \quad t \in (-\infty, 0] \end{aligned} \quad (2)$$

where $w_{\alpha 0}^{(\kappa)}(t)$ are prescribed initial accelerations.

The following theorem is based on Theorem 2.2.

THEOREM 3.1. Let the functions $r_{km}(t): [0,\infty) \to (0,\infty)$ satisfy the inequalities (k=1,2,...,N)

$$\frac{\sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{Q_0 P_0}{r_{km}^2(t)}}{1 - 3\sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{P_1}{r_{km}(t)}} < \frac{1 - \sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{M_2}{r_{km}(t)}}{3\sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{Q_0}{r_{km}^2(t)} \left[M_0 \left| t - \frac{r_{km}}{2c} \right|^2 + M_1 \left| t - \frac{r_{km}}{2c} \right| \right]}$$

then for every function $w_0 \in L_{loc}^\infty(R^1; R^1)$ satisfying the inequalities

$$\frac{\sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{Q_0 P_0}{r_{km}^2(t)}}{1 - 3P_1 \sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{1}{r_{km}(t)}} \leq w_0(t) <$$

$$< \frac{1 - \sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{1}{r_{km}(t)}}{3Q_0 \sum_{\substack{m=1 \\ m \neq k}}^{N} \frac{1}{r_{km}^2(t)} \left[M_0 \left| t - \frac{r_{km}}{2c} \right|^2 + M_1 \left| t - \frac{r_{km}}{2c} \right| \right]}$$

such that $|w_{\alpha 0}^{(m)}(t)| \leq w_0(t)$, there exists a solution of the initial value problem (2) belonging to $[L_{loc}^\infty(R^1)]^{3N}$ such that $r_{km}(t)$ are the distances between the particles for $t \geq 0$.

It is easy to see that

$$\lim_{t \to \infty} r_{km}(t) = \infty.$$

REFERENCES

1. SYNGE J.L. On the electrodynamic two-body problem of classical electrodynamics. Proc. Roy. Soc. (London), A, 177 (1940), 118-139.

2. PAULI W. Relativitatstheorie. "Encykl.d.Math. Wissenschften", Band V, Heft IV, Art. 19, 1921.

3. DRIVER R.D. A two-body problem of classical electrodynamics: the one dimensional case. Ann. Physics, v.21 (1968), 122-142.

4. DRIVER R.D. A neutral system with state dependent delay. J.Diff.Equations, v.54, N1 (1984), 73-86.

5. ANGELOV V.G. On the Synge equation in a three dimensional two-body problem of classical electrodynamics. J. Math. Anal. Appl., v.151 (1990) No2, 488-511.

6. DRIVER R.D., M.J.NORRIS A collinear n-body problem of classical electrodynamics. Nonlinear Phenomena in Math.Sci., Ed. V.Lakshmikantham, Academic Press, 1982, 329-334.

GRADIENT'S OPTIMIZATION METHOD AS THE BASE OF STRUCTURAL DESIGN OF A MULTIDIMENSIONAL ELECTROMECHANICAL SYSTEM

DANUTA JASIŃSKA-CHORCHAŃSKA
(NON-MEMBER OF IEEE)
Faculty of Precision Mechanics
CUPMiSP
Warsaw University of Technology
Chodkiewicza 8
02-525 Warsaw, POLAND

Abstract-The paper presents a method of optimization the constructional parameters considering maximization of the selected work's parameters of the multidimensional electromechanical system. The problem of optimization has been solved using Box-Wilson's method and the analysis of absolute function of sensitivity of the first order. This paper presents application of the above method and results of optimization.

I. INTRODUCTION

With the analysis of complex multidimensional objects, especially there where phenomena that are mechanic in their character are accompanied by magnetic and electromagnetic phenomena, the task of optimization an object on the basis of analysis of physical phenomena is very difficult, sometimes impossible. Most important problem for the electrical transducers for the measurement of displacement (for ex. for the inductive transducer) is optimization construction's parameters considering maximization of the metrological characteristics parameters. The Box-Wilson's method has been used to solve this problem [1]. The problem of optimization has been reduced to the classical problem of nonlinear programming with limitations (maximization output quantity y through changes of the input parameters $x_1, x_2, \ldots x_n$ in the preset intervals of permissible variations). Use has been made of the experimental design techniques, approximation of the object's properties and hypersurface of the I-st degree by the formula:

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n \qquad (1)$$

where $b_0, b_1, \ldots b_n$ are the coefficients of this relation in order to identify the model [1], [3].

The analysis of absolute function of sensitivity of the first order [2] by the formula:

$$S_{yx_j} = \left| \frac{\partial y}{\partial x_j} \right| \quad j=1,2,\ldots,n \qquad (2)$$

has been given information, which constructional parameters are more important than another in this optimization.

II. APPLICATION OF THE METHOD

The above method has been applied in the analysis of an inductive transducer system used for the measurement of displacements which works in the systems [3]:
1. of a hybrid transformer and
2. of a resistance bridge.

The measuring range has been taken as an output quantity y but it has been observed sensitivity y_1 (in mV/mm) and residual voltage y_2 (in V) this same time. The range of the linear characteristic's part in the separate experiments has been defined for the three values of the relative error (for 0,5%, 1% and 1,5%) based on the linear approximation [3].
The target of the analysis and experiments is obtaining of the maximal linear range (measuring range) shown in Fig. 1.

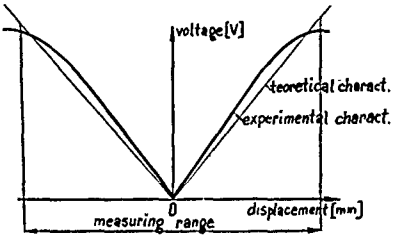


Fig. 1. Dependence of the output voltage's amplitude of the measuring system from the core's displacement

All parameters characterised the inductive transducer has been selected on the.
1. measurable input quantities which are the subject of the optimization process $(x_1, x_2, \ldots, x_n)$
2. and fixed measurable input quantities [3].

The constructional parameters regarded as the input quantities, have been presented in Table I.

TABLE I
MEASURABLE INPUT QUANTITIES

Genotation	The name of the parameter	Ground level	Permissible range of variations
x_1	Frequency of the power supply	10 kHz	0,5-20 kHz
x_2	Number of turns of each windings	1360	1÷5300 windings
x_3	Distance between the windings	2,5 mm	0,05÷38 mm
x_4	Length of the ferrite core	22,5 mm	1÷30 mm
x_5	Inside diameter of carcass	4,9 mm	3,4÷5,4 mm
x_6	Outside diam. of ferr. core	3,5 mm	3÷5 mm

Thus the problem is reduced to the maximization of y through the changes of the parameters $x_1, x_2, \ldots, x_6$ in the preset intervals of permissible variations for two systems of the transducer's work.

III. RESULTS

Owing to the realization of the experimental design technics (according to a two-level half design) 32 experiments have been made for each transducer's system of the work. The results have been presented in Table II and Table III [3]. Next experiments has been executed to finding the local maximum based on the Box-Wilson's method [3]. The local maximum has been found:
1. in the system of a hybrid transformer in 37 experiment (6,2 mm for 0,5%; 7,2 mm for 1%; 7,9 mm for 1,5%; sensitivity 220 mV/mm; residual voltage 0,015 V),
2. in the system of a resistance bridge in 36 experiment (7,7 mm for 0,5%; 8,7 mm for 1%; 9,3 mm for 1,5%; sensitivity 200 mV/mm; residual voltage 0,049 V).

That were the results of optimization. Identification of the coefficients of the form (1) basing according to the principle of the least sum square [3]. This coefficients shows Table IV and Table V.

TABLE IV
HYBRID TRANSFORMER

Error [%]	Coefficients					
	k_1	k_2	k_3	k_4	k_5	k_6
0,5	0,116	-0,078	-0,166	0,247	0,091	-0,028
1	0,159	-0,041	-0,141	0,334	0,078	-0,022
1,5	0,175	-0,025	-0,119	0,394	0,081	-0,019

TABLE V
RESISTANCE BRIDGE

Error [%]	Coefficients					
	k_1	k_2	k_3	k_4	k_5	k_6
0,5	-0,353	-0,147	-0,047	0,191	0,009	-0,009
1	-0,441	-0,078	-0,028	0,266	0,041	-0,041
1,5	-0,503	-0,078	-0,003	0,291	0,053	-0,022

Analsysed the absolute function of sensitivity of the first order (2) for system of a hybrid transformer, results has been got. This analysis informs, that outside diameter of the ferrite core is parameter, which makes up the least sensitivity, for system of a resistance bridge this analysis informs, that this same parameter makes up the least sensitivity.

IV. CONCLUSIONS

The gradient's optimization method has been appeared to be very effective to the optimization and structural design of the metrological properties of an inductive transducer for the measurement of displacements. The analysis of sensitivity can help in the selection of the more and less important constructional parameters to the next experiments and analysis. It seems that it can be successfully generalized for other systems especially for electromechanical measuring transducers.

REFERENCES

[1] Box G.E.P, Wilson K.B.: "On the experimental attaintment of optimum conditions. Journal of the Royal Statistical Society, Ser. B; No 1, 1951

[2] Frank P.M.: "Intrudocution of system sensitivity theory". Academic Press, New York 1978

[3] Jasińska-Choromańska D.: "Mathematical model of inductive transducer for measurement of displacement", doctor's dissertation, Warsaw University of Technology, Warsaw, 1983

TABLE II AND TABLE III [3]
RESULTS OF EXPERIMENTS OF A HYBRID TRANSFORMER AND OF A RESISTANCE BRIDGE

Factors x_n		x_1	x_2	x_3	x_4	x_5	x_6	$y_{0,5\%}$	$y_{1\%}$	$y_{1,5\%}$	y_4	y_5
Basic level x_{0n}		10	1360	2,5	22,5	4,9	3,5	5,0	5,8	6,3	234,8	0,015
Experimental step Δx_n		3	144	0,5	2,5	0,5	0,5					
Upper level (+)		13	1504	3	25	5,4	4					
Lower level (-)		7	1216	2	20	4,4	3					
Standardized variables t_c		t_1	t_2	t_3	t_4	t_5	t_6					
Experiment	1	+	+	+	+	+	+	6,1	7,1	7,8	231,3	0,065
	2	+	+	+	+	-	-	4,9	6,1	6,8	257,5	0,010
	3	+	+	+	-	-	+	4,4	5,1	6,0	288,2	0,020
	4	+	+	+	-	+	-	5,6	6,5	7,3	220,9	0,053
	5	+	+	-	-	+	+	5,2	6,1	6,6	239,6	0,031
	6	+	+	-	+	-	+	4,7	5,7	6,3	270,1	0,004
	7	+	+	-	+	+	-	5,9	6,8	7,4	216,0	0,051
	8	+	+	-	-	-	-	4,1	5,0	5,6	278,4	0,016
	9	+	-	-	+	+	+	5,2	6,2	6,9	235,8	0,031
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	38	13,0	1294	2,20	26,7	5,08	3,38	6,2	7,0	7,5	216,5	0,013

NEW METHODS AND RESULTS IN 3D-DEVICE-SIMULATION

H. Elschner[a], W. Klix[a], R. Dittmann[a],
R. Reuter[b], R. Stenzel[c]
(a) University of Technology Dresden
Institute of Electrotechnics/Electronics
Mommsenstr. 13, D-8027 Dresden
(b) IBM Scientific Center Heidelberg
Tiergartenstr. 15, D-6900 Heidelberg
(c) University of Transport and Communication
Institute of Electrotechnics
Friedrich-List-Platz 1, D-8010 Dresden

Abstract In this paper we present efficient methods for the solution of large systems of linear equations arising within the 3D-Device Simulator SIMBA, and simulation results for trench cells, IGBT's and an IPG-FET.

1. Introduction

We describe numerical methods, which are now implemented in the 3D-Device Simulator SIMBA, and results of the simulation of new semiconductor devices. Three dimensional simulations become very important for the development and optimization of modern microelectronic devices and VLSI-ICs.
The first release of SIMBA was available for industrial applications in 1982. An analytical model and an interface to the Process Simulator DUPSIM are included in SIMBA for nonplanar surfaces and the doping profile.
The behavior of the semiconductor device (Si, GaAs, InP) is modelled by the well-known system of semiconductor equations:

$$\mathrm{div}(\varepsilon * \mathrm{grad}\,\psi) = -(p - n + CN) \tag{1}$$

$$\mathrm{div}\vec{S}_p = -q * (R - G + \frac{\partial p}{\partial t}) \tag{2}$$

$$\mathrm{div}\vec{S}_n = q * (R - G + \frac{\partial n}{\partial t}) \tag{3}$$

$$\vec{S}_p = -q * (\mu_p * p * \vec{E} + D_p * \mathrm{grad}\,p) \tag{4}$$

$$\vec{S}_n = -q * (\mu_n * n * \vec{E} - D_n * \mathrm{grad}\,n) \tag{5}$$

By using the band parameter model [1], it is possible to simulate hetero devices. The recombination model includes the SHOCKLEY/READ/HALL recombination with doping dependent life times and the Auger recombination. A modified model, according to CHAUGLY/THOMAS is used for the carrier mobility [2].

2. Numerical Methods

The set of equations (1) (5) is transformed with the GAUSSian law and then discretized by a seven point difference scheme in a rectangular nonuniform grid. The backward EULER formula, combined with a predictor, is used for the time discretization.
To describe the current densities, the GUMMEL/SCHARFETTER method is used. The base cell for the discretization is a cube with an inner grid point. At the outer planes of the device some parts of the cell are removed and the grid point is located at the surface of the device. It is possible to describe a nonplanar surface by means of this technique.
The Poisson and the continuity equations are solved consecutively according to GUMMEL's method. Normally we use for the discretized form of the Poisson equation $\rho = p(\psi), n = n(\psi)$ and solve it nonlinear by using NEWTON's method. Sometimes it is useful to consider p,n in (1) as independent on ψ. Then a system of linear equations results. The discretized continuity equations are treated linearly. The set of unknowns consists of the electrostatic potential and the carrier densities. The large systems of linear equations

$$A * x = b$$

are solved with a modified ICCG method [3] and the ICBCG method [4],[5] for the Poisson equation and the continuity equations, respectively. In principle the algorithms read as follows.
- initialization

$$r_0 = b - A * x_0 \;;\; rs_0 = r_0$$

$$p_0 = (L.U)^{-1} * r_0 \;;\; ps_0 = (\tilde{L}.\tilde{U})^{-1} * rs_0$$

- loop until convergence

$$\alpha = \frac{rs_i * r_i}{ps_i * (A * p_i)} \qquad dls = (\tilde{L}.\tilde{U})^{-1} * rs_{i+1}$$

$$r_{i+1} = r_i - \alpha * (A * p_i) \qquad \beta = \frac{dls * (A^T * ps_i)}{p_i * (A^T * ps_i)}$$

$$dl = (L.U)^{-1} * r_{i+1} \qquad p_{i+1} = dl + \beta * p_i$$

$$x_{i+1} = x_i + \alpha * p_i \qquad ps_{i+1} = dls + \beta * ps_i$$

$$rs_{i+1} = rs_i - \alpha * (A^T * ps_i)$$

where $(L.U)$ and $(\tilde{L}.\tilde{U})$ denote incomplete decompositions of the matrices A and A^T, respectively.
SIMBA runs on VAX/VMS, IBM 3090, CONVEX, and SUN-workstations. Essential parts of the code are vectorized, which results in a remarkable decrease in CPU time (a factor of ca. 4 was measured for typical examples). It is planned to study the possibilities of further improvements in order to exploit the vector and parallel capabilities of the IBM 3090 Multiprocessor. As the solution of the systems of linear equations is the most time consuming part (about 80 % of CPU time) it will be the favorite candidate for parallelization. Some of possible techniques for the parallelization of the above stated algorithm may be:

- spread the simple vector operations ($y = z + \alpha * x$ and $y = z * x$) across several processors,
- in a matrix-vector-product partition the matrix horizontally and compute each corresponding part of the resulting vector separately in one processor,
- apply block techniques in order to get a parallel solver for $(L.U)^{-1} * x$ and $(\tilde{L}.\tilde{U})^{-1} * x$.

3. Results

The punch-through behavior of trench cells in 2D- and 3D case is a first result of the simulations performed with SIMBA. The sketch of the device and the current voltage relations are shown in Fig. 1 and 2. The difference between the two calculated punch-through voltages is about 2 Volts, caused by the influence of the back area in 3D case. About 200 000 grid points were used for this simulation.
The structure of a n-channel IGBT (insulated gate bipolar transistor) is shown in Fig. 3. The thickness of the gate oxid is 300 nm. The IGBT was biased to a gate voltage of 10 V to calculate the output characteristic, shown in Fig. 4.
By means of SIMBA, it was possible to simulate so called IN-Plane-Gated (IPG) field effect transistors. WIECK and PLOOG developed these devices [6], which have a quasi one-dimensional (Q1D) tunable carrier channel. Fig. 5 shows the structure, which was used for the 3D simulation. The hetero-structure consists of an undoped GaAs layer, a 20 nm undoped AlGaAs spacer, and a 50 nm doped AlGaAs layer. The channel is insulated laterally from the two dimensional electron systems (2DES), which serve as gates. The conductivity in the Q1D channel is controlled via a gate to-source voltage V_{GS}, which is applied between the channel and adjacent 2DES regimes. Because the regions near the insulated paths have lower carrier densities and mobilities, they were included in our simulation. The calculated transfer characteristic at room temperature is represented in Fig. 6, which shows a threshold voltage of about 7 V and a drain current of 11.6 µA at $V_{GS} = 0$. The results of the simulation agree well with those obtained experimentally in [5].

4 Conclusions

- The ICBCG method is an efficient algorithm for the solution of large linear systems arising in device simulation.
- The use of a vector processor accelerates the overall performance of the simulation runs by a factor of ca 4.
- The simulation of trench cells and an IPG-FET compares very well with experimental results obtained by other authors.

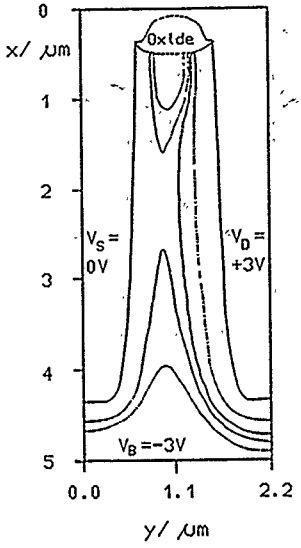


Fig 1. Sketch of the simulated trench cells with equipotential lines

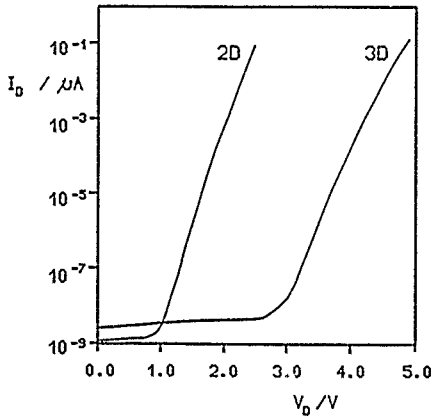


Fig.2. Punch-through behavior in 2D- and 3D-case

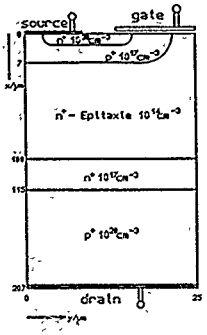


Fig 3 Structure sketch of the IGBT

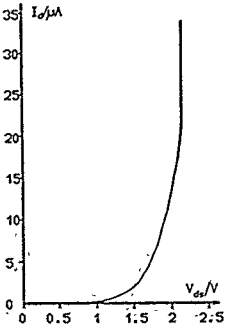


Fig 4 Output characteristic of the IGBT

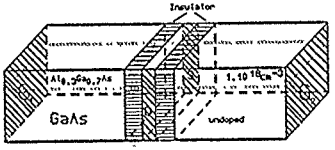


Fig 5. Structure sketch of the IPG field effect transistor

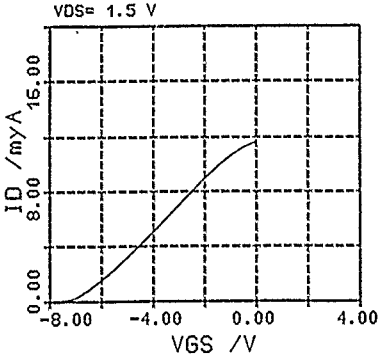


Fig 6. Calculated Transfer characteristic of the IPG field effect transistor

5. References

[1] Yokoyama, K., et al., IEEE Trans. on ED, VOL-ED 30 (1983), pp. 1283-1288

[2] Klix, W., 4. Tagung Schaltkreisentwurf, Technische Universität Dresden, 1990, pp. 38 42

[3] Wada, T., et al., Electronic letters, Vol. 18, No. 6, 1982, p. 265

[4] Voigt, R., Belegarbeit Ingenieurpraktikum, Technische Universität Dresden, Sektion Informationstechnik 1989

[5] Elschner, H., et al., Proc. of NASECODE VI Conference, Dublin 1989, pp. 148-153

[6] Wieck, A D., Ploog, K., Appl. Phys. Lett., 56 (1990), p. 928

SIMULATION AND OPTIMIZATION OF CIRCUITS BY MEANS OF WAVEFORM RELAXATION METHODS

KLAUS R. SCHNEIDER
Kœl-Weierstraß-Institute of Mathematics
Mohrenstraße 39
Berlin, O - 1086, F.R.G.

Abstract - The waveform relaxation method (WRM) is an important tool to compute the solutions of Cauchy problems for very large systems of differential-algebraic systems. By means of Pontrjagin's maximum principle, problems of optimal control can be reduced to the solution of boundary value problems. In case of large structured systems it is efficient to compute the solution of the corresponding boundary value problem by WRM. We give conditions for the convergence (both global and quadratically local) of the WRM.

1. GLOBALLY CONVERGENT WRM

The transient analysis of electronic circuits is equivalent to the solution of the Cauchy problem for the differential-algebraic system

$$\frac{d\xi}{dt} = \tilde{f}(\xi, d\xi/dt, \eta, t), \quad \eta = \tilde{g}(\xi, d\xi/dt, \eta, t) \tag{1}$$
$$\xi(t_0) = \xi_0, \; t \in S := (t_0, t_0 + T).$$

In case $dim\ \xi + dim\ \eta \gg 1$ circuit simulation is a time consuming task. In overcoming this difficulty WRM plays a central role, it is an iterative method for determining the solution of (1) which is based on the decomposition of the large system into subsystems and on the independent (effective) solution of each subsystem, WRM is immedeately suitable for parallel techniques. The corresponding canonical scheme reads [2]

$$\begin{aligned} \frac{dx^k}{dt} &= f(x^k, x^{k-1}, dx^{k-1}/dt, z^{k-1}, t), \\ z^k &= g(x^k, x^{k-1}, dx^{k-1}/dt, z^{k-1}, t) \\ x^k(t_0) &= \xi_0, \; t \in S, \; k = 1, 2, \ldots \end{aligned} \tag{2}$$

Concerning f and g we suppose

(V). $f : R^n \times R^n \times R^n \times R^m \times R \to R^n$, $g : R^n \times R^n \times R^n \times R^m \times R \to R^m$ are continuous and satisfy $\forall\, w, \bar{w} \in R^{3n+m}$ *where* $w := (x_1, x_2, x_3, z), \bar{w} := (\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{z}), \; t \in R$

$$|f(w,t) - f(\bar{w},t)| \le \sum_{i=1}^{3} \kappa_i |x_i - \bar{x}_i| + \kappa_4 |z - \bar{z}|$$
$$|g(w,t) - g(\bar{w},t)| \le \sum_{i=1}^{3} \lambda_i |x_i - \bar{x}_i| + \lambda_4 |z - \bar{z}|$$

where the spectral radius $\sigma(L)$ of the matrix

$$L := \begin{pmatrix} \lambda_3 & \lambda_4 \\ \kappa_3 & \kappa_4 \end{pmatrix}$$

obers $\sigma(L) < 1$.

The proof of the following theorem can be found in [3].

THEOREM 1.1. *Assume hypothesis (V) holds. Then the Cauchy problem (1) has to any $T > 0$ a unique solution and the iteration scheme (2) converges for any initial guess to the unique solution.*

2. QUADRATICALLY CONVERGENT WRM

In an appropriate Banach space we introduce the operator

$$\begin{aligned} \tilde{F}(w)(t) := (\xi(t) &- \xi_0 + \int_{t_0}^{t} \tilde{f}(\xi(s), d\xi/ds, \eta(s), s)ds, \\ z(t) &- \tilde{g}(\xi(t), d\xi/dt, \eta(t), t). \end{aligned}$$

Then, the Cauchy problem (1) is equivalent to

$$\tilde{F}(w) = 0 \tag{3}$$

In general, the operator $\tilde{F}$ has a special structure which can be exploited for a block iterative scheme. In the simplest case we represent (3) in the form

$$\begin{aligned} F_1(w_1, w_2) &= 0 \\ F_2(w_1, w_2) &= 0 \end{aligned} \tag{4}$$

Concerning the system (4) we assume

(V_1). (4) has a unique solution w^*.
(V_2). $\tilde{F}$ is twice continuously Fréchet diffentiable near w^*.
(V_3).

$$F_{1,w_1}, \quad \begin{pmatrix} F_{1,w_1} & F_{1,w_2} \\ F_{2,w_1} & F_{2,w_2} \end{pmatrix}$$

have a bounded inverse at $w = w^*$.

Consider the block iteration scheme

$$\begin{aligned} w_1^{k+1} &= w_1^k + F_{1,w_1}^{-1}(w_1^k, w_2^k) F_1(w_1^k, w_2^k) \\ w_2^{k+1} &= w_2^k + [F_{2,w_2}(w_1^{k+1}, w_2^k) - \\ &- F_{2,w_1}(w_1^{k+1}, w_2^k) F_{1,w_1}^{-1}(w_1^k, w_2^k) F_{1,w_2}(w_1^{k+1}, w_2^k)]^{-1} F_2(w_1^{k+1}, w_2^k) \end{aligned} \tag{5}$$

THEOREM 2.1. *Assume the hypotheses (V_1) – (V_3) hold. Then the iteration scheme (5) converges with 1-step Q-order 2 to the unique solution w^* of (4) provided the initial guess w^0 is sufficiently near w^*.*

The iteration scheme (5) is related to a similar procedure due to Hoyer and Schmidt [1] in case of finite dimensional space. The general case will be treated in [4].

3. OPTIMIZATION PROBLEMS FOR LARGE CIRCUITS

The problem to minimize the functional

$$J[u] = \int_0^T h(x(t), t, u(t))dt$$

where the control function u belongs to some set U and x satisfies

$$\frac{dx}{dt} = f(x, t, u), \quad x(0) = x_0,$$

can be reduced by Pontrjagin's maximum principle to the boundary value problem

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H}{\partial \lambda}(x, \lambda, t), \\ \frac{d\lambda}{dt} &= -\frac{\partial H}{\partial x}(x, \lambda, t), \\ x(0) &= x_0, \lambda(0) = 0. \end{aligned} \tag{6}$$

where H is the corresponding Hamiltonfunction.

In case $dim\, x = dim\, \lambda \gg 1$ we may in general assume that H has a special structure. This suggests the use of block diagonal procedures for the numerical solution of (6), especially WRM. The corresponding iteration scheme, related immedeately to (6), can be written in the form

$$\begin{aligned} x^k(t) &= x^0 + \int_0^T h_1(x^k, x^{k-1}, \lambda^k, \lambda^{k-1}, s)ds \\ &\quad + \int_T^t h_1(x^k, x^{k-1}, \lambda^k, \lambda^{k-1}, s)ds, \\ \lambda^k(t) &= \int_0^T h_2(x^k, x^{k-1}, \lambda^k, \lambda^{k-1}, s)ds \\ &\quad + \int_0^t h_2(x^k, x^{k-1}, \lambda^k, \lambda^{k-1}, s)ds. \end{aligned} \tag{7}$$

With respect to the functions h_1 and h_2 we assume that there are positive numbers $k_i^{(j)}$, $l_i^{(j)}$; i,j = 1,2, such that $\forall\, x_1, x_2, \lambda_1, \lambda_2, \bar{x}_1, \bar{x}_2, \bar{\lambda}_1, \bar{\lambda}_2 \in R^n$, $t \in R$
(H).

$$\begin{aligned} |f_j(x_1, x_2, \lambda_1, \lambda_2, t) - f_j(\bar{x}_1, \bar{x}_2, \bar{\lambda}_1, \bar{\lambda}_2, t)| \leq \\ \leq \sum_{i=1}^{2} (k_i^{(j)}|x_i - \bar{x}_i| + l_i^{(j)}|\lambda_i - \bar{\lambda}_i|). \end{aligned} \tag{8}$$

Furthermore we introduce the matrices

$$M_i := \begin{pmatrix} k_i^{(1)} & l_i^{(1)} \\ k_i^{(2)} & l_i^{(2)} \end{pmatrix}, \quad i = 1, 2.$$

Now we are ready to formulate the following theorem.

THEOREM 3.1. *Assume the hypothesis (H) holds. Additionally we assume*

$$To((I - TM_1)^{-1}M_2) < 1. \tag{9}$$

Then the iteration scheme (7) converges to the unique solution of (6) for any initial guess.

Theorem 3.1 can be interpreteedd as follows: Under the given conditions the successive optimization of the subsystems leads to the optimal control of the full system. In case that (9) cannot be fulfilled we may apply the approach described in section 2.

REFERENCES

[1] W. Hoyer, and J.W. Schmidt. Newton-type decomposition methods for equations arising in network analysis. Z. Angew. Math. Mech. 64, 397 - 405 (1984).

[2] E. Lelarasmee, A.E. Ruehli, A.L. Sangiovanni-Vincentelli, Waveform relaxation method for the time-domain analysis of large scale integrated circuits. IEEE Trans. Computer-aided Design, CAD-1, 131-145 (1982).

[3] K.R. Schneider, A remark on the wareform relaxation method. Int. J. Circuit Theory Appl. 18 (1990).

[4] K.R. Schneider Quadratically convergeat wareform relaxation schemes. (in preparation).

ON EQUIVALENT INDEX ONE FORMULATIONS FOR HIGHER-INDEX DIFFERENTIAL-ALGEBRAIC EQUATIONS

Sebastian Reich
Karl-Weierstraß-Institute of Mathematics
Mohrenstraße 39
Berlin, O-1086, F.R.G.

Abstract Mathematical modeling of lumped physical systems such as electronic circuits and mechanical systems leads frequently to higher-index differential-algebraic equations (DAEs). In many cases a computational analysis of these DAEs is of interest. However, the well-known numerical methods, like BDF or Runge-Kutta, are applicable in general to index one problems only. In this paper we show that any higher-index DAE can be transformed to an index one DAE such that the set of solutions is kept invariant during the transformation.

1 INTRODUCTION

DAEs are frequently identified as implicit equations

$$F(t,x,x') = 0, \qquad F:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n \qquad (1)$$

for which x' cannot be expressed explicitly as a function of t and x. Such DAEs arise in many areas of science and engineering. In particular mechanical systems and electronic circuits may be modeled using equations of the type (1).

The problem (1) is structural quite different from an ODE, and existing ODE methods are in general not applicable to DAEs without extensive modifications. The variation in the structure of (1) from an ODE is usually quantified by an integer called the index of the problem [1]. While index zero and index one problems can be solved by the well-known methods like BDF or Runge-Kutta, it is difficult to solve higher-index problems numerically. For that reason several techniques for the reduction of higher-index DAEs to DAEs with a lower index are discussed in the literature [1],[2],[3],[4]. However, most of these techniques introduce additional degrees of freedom during the reduction. This results in certain integral invariants for the reduced DAE which the numerical method may not keep constant during integration.

In this paper we discuss a technique for the reduction of higher-index problems to DAEs of index one that do not introduce any additional degrees of freedom, integral invariants respectively. To distinguish reduction techniques with this property from arbitrary reductions, we call them, as suggested in [2], index transformations.

2 MATHEMATICAL BACKGROUND

To simplifiy the notations, we rewrite (1) as a time-invariant DAE

$$F(z,x,x') = 0$$
$$z' = 1$$

and consider DAEs of the type

$$F(x,x') = 0, \qquad F:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n \qquad (2)$$

only. We associate with DAEs of this type for any non-negative integer m, as long as the involved mappings are defined and differentiable, the nonlinear system of equations

$$\begin{aligned} F_0(x,p) &= 0 \\ &\vdots \\ F_m(x,p) &= 0 \end{aligned} \qquad (3)$$

with

$$F_0(x,p) := F(x,p),$$

$$F_{k+1}(x,p) := Q_k(x,p)D_xF_k(x,p)p \quad (k=0,\ldots,m-1)$$

where

$$Q_k(x,p)$$

is a projection along

$$\mathrm{im}(D_pF_k(x,p))$$

and

$$D_xF_k(x,p),\ D_pF_k(x,p)$$

denote the partial derivatives with respect to x and p. We call (3) the derivative array of order m, and we denote the solution set of (3) by $L(m) \subset \mathbb{R}^n\times\mathbb{R}^n$ and the projection of $L(m)$ onto the first component by $M(m) \subset \mathbb{R}^n$.

Let us assume now that there exists a non-negative integer i such that

- C1: $M(i)$ is a differentiable manifold.
- C2: for any $x \in M(i)$ there exists a unique $p \in \mathbb{R}^n$ with $(x,p) \in L(i)$.

If such an integer exists, then the smallest integer, that satisfies the conditions C1 and C2, is called the index of the problem [5].

Let (2) be a DAE of index 1, then the condition C2 defines a unique mapping $v:M(1)\to\mathbb{R}^n$ which is a vector field on $M(1)$. We call the manifold $M = M(1)$ the configuration space and the mapping v the corresponding vector field of the problem Obviously, the solutions of the corresponding vector field are identical with the solutions of the given DAE. Thus, in this case, the well-known existence and uniqueness results for vector fields are applicable to DAEs [5].

In several papers [1],[3], it is assumed that the derivative arrayof order i can be rewritten in the form

$$F^1(x,p) = 0, \quad F^1: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \qquad (4.1)$$

$$F^2(x) = 0, \quad F^2: \mathbb{R}^n \to \mathbb{R}^m \qquad (4.2)$$

with

$$\operatorname{rank}(D_p F^1(x,p)) = n$$

and

$$\operatorname{rank}(D_x F^2(x)) = m.$$

Thus (4.2) yields the configuration space M of the problem and (4.1) can be considered as the implicit description of an ODE

$$x' = g(x), \quad g: U \to \mathbb{R}^n \qquad (5)$$

which is called the underlying ODE of the problem. Clearly, the restriction of the mapping g to the configuration space M is identical to the corresponding vector field v. Thus the configuration space M is an invariant manifold of the underlying ODE (5).

3 INDEX TRANSFORMATIONS

For the computational analysis of higher index problems it seems reasonable to integrate the underlying ODE instead. However, most numerical methods do not keep the configuration space M invariant during the integration. This results in the well-known numerical drift-off. Therefore we derive lower index formulations which maintain the configuration space of the problem in this chapter.

First we generalize an index transformation technique which was proposed by GEAR in [3]. Let us consider the formulation (4) of the derivative array (3) which is an overdetermined system of equations. Then the generalization of the stabilization technique, as proposed by GEAR, results in the DAE

$$\begin{aligned} F^1(x,x' - D_x F^2(x)^T \mu) &= 0 \\ F^2(x) &= 0 \end{aligned} \qquad (6)$$

where $\mu \in \mathbb{R}^m$ and ^{T} denotes the transposed of a linear mapping. This DAE is a DAE of index 2 and has exactly the same solutions as (2) with respect to the variable x. Furthermore, we have $\mu = 0$ for any solution of (6).

In many cases it might be difficult or even impossible to reformulate the derivative array (3) in the form (4). Therefore we propose here an index transformation technique which does not require the formulation (4).

For any (x,p) let us consider the linear subspace

$$N(x,p) := \ker \begin{bmatrix} Q_0(x,p) D_x F_0(x,p) \\ \vdots \\ Q_1(x,p) D_x F_1(x,p) \end{bmatrix}.$$

If $\dim(N(x,p)) = \text{const.}$ for all (x,p), then N(x,p) does not depend on the variable p and we have

$$T_x M(1) = N(x)$$

where

$$T_x M(1)$$

denotes the tangent space of M(1) at x [5]. Let P(x) be a projection onto N(x). We replace now x' in (2) by P(x)x' and obtain the new DAE

$$F(x, P(x)x') = 0. \qquad (7)$$

This substitution makes sense because any solution of (2) has to satisfy

$$x' \in T_x M(1)$$

and thus

$$x' = P(x)x'.$$

Consequently we can formulate the following

Theorem: The DAE (7) is a DAE of index one and has exactly the same solutions as the given DAE (2) of index 1.

Proof: Let $L(m)^\#$, $M(m)^\#$ be the sets associated with the derivative array of order m corresponding to the DAE (7). Then, by definition of the DAE (7), we obtain that

$$(x,p) \in L(0)^\#$$

iff

$$(x,z) \in L(0) \text{ and } z = P(x)p \qquad (8)$$

However, because [5]

$$(x,p) \in L(1)$$

iff

$$(x,z) \in L(0) \text{ and } z \in T_x M(1),$$

we rewrite the condition (8) to

$$(x,z) \in L(1) \text{ and } z = P(x)p.$$

This implies that

$$M(0)^\# = M(1)$$

and

$$L(1) = L(0)^\# \cap TM(0)^\#$$

which is equivalent to [5]

$$L(1) = L(1)^\#. \qquad \blacksquare$$

References

[1] Brenan, K.E., Campbell, S.L.; Petzold, L.R.: The Numerical Solution of Initial Value Problems in DAEs. North Holland Publ. Co., 1989

[2] Eich, E.; Führer, C., Leimkuhler, B., Reich, S.: Stabilization and Projection Methods for Multibody Dynamics. Techn. Univ. Helsinki, Report A 281, 1990

[3] Gear, C.W.: DAEs Index Transformations. SIAM JSSC, 9(1988)1, pp. 39-47

[4] Reich, S.: On a Geometrical Interpretation of DAEs. Circuits, Systems; Signal Processing, 9(1990)4, pp. 367-382

[5] Reich, S.: Existence and Uniqeness Results for DAEs. to be published in: Lecture Notes in Mathematics, eds.: März, R.; Hanke, M.; Griepentrog, E.

CONVERGENCE ACCELERATION FOR THE BICONJUGATE GRADIENT METHOD

WOLFGANG BURMEISTER
Dresden University of Technology
Department of Mathematics
Zellescher Weg 12-14
D-O-8027 Dresden , Germany

Abstract Solving large-scale linear systems of equations by conjugate gradient (CG) or related methods requires a good preconditioning of the original coefficient matrix to obtain an approximate solution in a reasonable number of steps. In many applications (higher order difference schemes, singularly perturbed problems) the matrix is not symmetric and the standard CG method not applicable.
In this case, the so-called biconjugate gradient method will be interpreted as an ordinary CG method applied to an extended system with symmetric but indefinite matrix. This idea can be used to derive other methods.

I. INTRODUCTION

Consider the linear system of equations

$$Ax = b \qquad (1)$$

with symmetric matrix A. Let a (possibly indefinite) scalar product

$$(u,v) = u^T C v$$

be defined by another symmetric matrix C. Then the conjugate gradient method for this system with respect to this scalar product is given by

x_1 starting vector , $r_1 = b - Ax_1$, $p_1 = Cr_1$

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k p_k \\ r_{k+1} &= r_k - \alpha_k A p_k \\ p_{k+1} &= Cr_{k+1} + \beta_k p_k \end{aligned} \qquad (2)$$

with

$$\alpha_k = \frac{r_k^T C r_k}{p_k^T A p_k} \quad \text{and} \quad \beta_k = \frac{r_{k+1}^T C r_{k+1}}{r_k^T C r_k} .$$

The choice of C affects the convergence rate of the method. It is most natural to take as C an approximation to the inverse A^{-1} by means of incomplete Cholesky factorization.
The method terminates when $r_k = 0$; then $x = x_k$ is a solution of (1). However, it breaks down whenever $p_k^T A p_k = 0$ or $r_k^T C r_k = 0$. In this situation the process is repeated with another starting vector x_1.

II. THE BCG METHOD

Assume now a system (1) with a not necessarily symmetric matrix A together with the transposed system

$$A^T y = c \ .$$

Both equations can be collected together as

$$\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix} \qquad (3)$$

and solved by the above method with A, C replaced by $\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix}$ where the new C is again an approximation to A^{-1} obtained from an incomplete LU factorization of A.
The resulting process computes iteratively three sequences of vectors $\begin{pmatrix} x_k \\ y_k \end{pmatrix}$, $\begin{pmatrix} r_k \\ s_k \end{pmatrix}$ and $\begin{pmatrix} p_k \\ q_k \end{pmatrix}$ according to

x_1, y_1 starting vectors, $r_1 = b - Ax_1$,

$$\begin{aligned} s_1 &= c - A^T y_1 \ , \ p_1 = Cr_1 \ , \ q_1 = C^T s_1 \\ x_{k+1} &= x_k + \alpha_k p_k \\ y_{k+1} &= y_k + \alpha_k q_k \end{aligned} \qquad (4)$$

$$r_{k+1} = r_k - \alpha_k A p_k \ , \quad p_{k+1} = Cr_{k+1} + \beta_k p_k$$

$$s_{k+1} = s_k - \alpha_k A^T q_k \ , \quad q_{k+1} = C^T s_{k+1} + \beta_k q_k$$

with

$$\alpha_k = \frac{s_k^T C r_k}{q_k^T A p_k} \quad \text{and} \quad \beta_k = \frac{s_{k+1}^T C r_{k+1}}{s_k^T C r_k} .$$

This method (with C=I) is known as the biconjugate gradient (BCG) method. The present approach now suggests to study other choices in (2) for solving the same system (3).

III. ANOTHER SPECIAL CASE OF METHOD (2)

Again A in (2) is replaced by $\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$ in order to solve (3), but C is taken as $\begin{pmatrix} F^T F & 0 \\ 0 & EE^T \end{pmatrix}^{-1}$ where $A \approx EF$ is an approximate factorization of A with easily invertible factors E and F. In this case the updating of x_k, y_k, r_k, s_k proceeds as in (4), but

$$p_1 = F^{-1}F^{-T}s_1 \ , \ p_{k+1} = F^{-1}F^{-T}s_{k+1} + \beta_k p_k$$

and

$$q_1 = E^{-T}E^{-1}r_1 \ , \ q_{k+1} = E^{-T}E^{-1}r_{k+1} + \beta_k q_k$$

with

$$\alpha_k = \frac{\gamma_k}{2 q_k^T A p_k} \ , \ \beta_k = \frac{\gamma_{k+1}}{\gamma_k}$$

$$\gamma_k = s_k^T F^{-1}F^{-T}s_k + r_k^T E^{-T}E^{-1}r_k \ .$$

REFERENCE

R.Fletcher, Conjugate Gradient Methods for Indefinite Systems, in: Lect.Notes in Math. 506, Springer V. 1976, pp.73-89.

CIRCUIT MODELING OF POWER ELECTRONICS DEVICES

Henry Güldner
Hochschule für Verkehrswesen
Lehrstuhl Leistungselektronik
Friedrich-List-Platz 1
O-8010 Dresden
Germany

Frank Dahms
Siemens AG
Bereich Verkehrstechnik
Werner-von-Siemensstraße 67
W-8520 Erlangen
Germany

Abstract - According to the special feature of power electronics systems a simulation program has been developed and tested providing a separate modeling of the sub-components on an interdisciplinary basis.

While the ac supply voltages (1), the power electronics circuit (2) and load (3) are primarily described on a network model basis, the modeling of control units (5) is made with block-oriented functional modules, whereas for modeling the gating circuit (4) a control graph is used (Fig. 1).

The analysis program is an integral part of a design system equipped with a data base which can be installed in UNIX compatible operating systems. Through its implementation in the programming language C the demand for a good portability is fulfilled. As regards analytical results required for design purposes which are not directly computed, a postprocessor with graphic representation is provided.

I. INTRODUCTION

The design of power electronics systems is an iterative process in which the satisfaction of the function, being on its turn determined by the automation process aimed at, is linked with the requirement of harmonizing the permissible static and dynamic stress limits for active and passive components. For design evaluation purposes an analysis model is necessary which does not only reflect the system-relevant properties but is also compatible with engineering aspects. In order to include into modeling the interaction between the switching behaviour of the power electronics devices (thyristor, transistor, GTO, IGBT) and the actual system behaviour (load) it is necessary to make use of the analytical methods designed for microelectronic circuits.

II. HYBRID SYSTEM MODELING AND ANALYSIS

In Fig. 1 the structure of a power electronics system is shown.

Electric circuits are usually designed on a network basis while control systems, owing to their mainly functional aspects, are not currently designed on the level of their electrical networks.

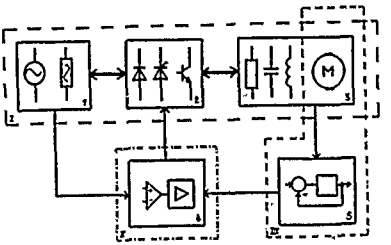


Fig. 1: Structure of a power electronics system

A. Model Properties

The network model of an electric dynamic circuit is a mixture of algebraic and differential equations. It is non-linear especially if continuous models for power electronics devices are included. Moreover these models are responsible for the rigidity of the differential equations caused by the dynamic components of the model (input/output behaviour).

The network equation system is set up on the basis of the Kirchhoff laws, the current/voltage functions of the elementary networks and circuit topology information according to the modified modal approach. Owing to the possibility that there is a choice of representing the two-pole relations either in an admittance or impedance form this hybrid network description is especially suitable for realizing a model of an "ideal switch".

The differential equation system which can be read from the block structures is treated according to the Runge-Kutta method of the fourth order due to its insensitivity to discontinuities in the behaviour of the system parameters. The block computation sequence is fixed according to the signal direction by means of the tree sorting method. The causality inherent in the gating method of power electronics systems is noted down in the form of state-event sequences. The interpretation of steady state graphs in the case of changing states gives rise to a modification of the steplength of integration.

B. Numerical aspects

Considering the special properties of the algebro-differential equation system (network) it is solved step by step by means of the Newton-Raphson iteration. The energy storage equations are discretized on a two-terminal network by BDF (Backward Differentiation Formulas) or alternatively by the trapezoid formula.

For iteration the linearized equation systems are repeatedly solved.

An improved convergence behaviour of the Newton method with regard to the switching point can be reached by means of a defect-reducing embedding of the network equation system.

A special damping of the Newton correction vector makes sure that the order of the iterated does not change if the behaviour of the function is very flat. The break-off limit of the Newton iteration is adapted to the momentary order of the unknown in the equation system. To maintain the stability of BDF which is controllable with regard to its steplength and order, order-dependent boundaries of the steplength-changing factor are necessary.

Owing to the possibility of changing the order of integration formulas, the integration method can be started by itself, and moreover there is also the possibility of suppressing the influence of the solutions obtained before a switching point on the solutions immediately after a switching process. For this purpose the trapezoid formula is combined with the A-stable implicit Euler formula.

For event inclusion methods based on polynom interpolations have been tested with a view to convergence acceleration. The inverse Hermitian interpolation is combined with the Regula falsi to fulfil the conditions made with regard to the derivation values of the polynom. As the event includes an increment corresponding to the minimum permissible steplength, it is necessary to change from the interpolation method to the robust interval halving method.

C. Results and Prospects

The results obtained from the analysis of the motor speed control for a DC drive are shown in fig. 2. As can be seen the armature circuit of the DC drive is fed by a three-phase full-wave bridge.

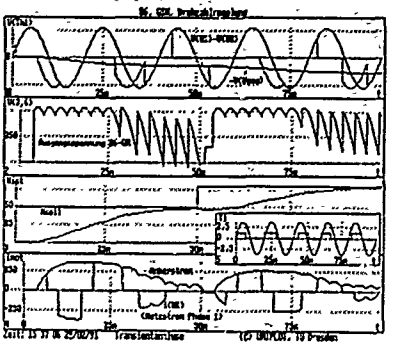


Fig. 2: Results obtained from the analysis of the motor speed control for a DC drive

Besides the design of refined valve models future developments should aim at using numerical optimization methods for design purposes. The computation of the gradient of objective functions is possible by means of a sensitivity analysis which is applicable to the network model. The LU decomposition of the Jacobi matrix during the Newton iteration is the prerequisite for the efficiency of this analysis which is made simultaneously with the transient analysis. At each discretization point a linear equation system has to be solved per design parameter.

Numerical solution of partial differential equations with a time-dependent region used in process simulation as demonstrated by means of epitaxy

Dirk Bothmann
University of Technology Dresden
Institute of Electrotechnics/Electronics
Mommsenstr. 13, D-8027 Dresden, Germany

1. Introduction

The production of semiconductor structures on a silicon wafer will be simulated by process simulation. The redistribution of impurities in the silicon wafer, called dopant diffusion, and changes in the structure of various layers, called structuring processes, are most important during production.
In this case, a silicon wafer is considered, with epitaxy being regarded as a special step. This production step is done in a reactor, in which the silicon wafer is placed and a gas stream transports both, silicon compounds and dopants. In this process cristalline silicon is deposited, so that the substrat region for dopant diffusion changes in time. There is also a impurity redistribution in the wafer by diffusion due to the high temperature, which is coupled with a dopant flow through the silicon/gas surface because of the dopants in the reactor gas. The present paper discusses a programme system simulating epitaxy in a two-dimensional case (as well as other technological steps) in the context of changing layers and diffusion, including modelling of the boundary conditions.

2. Time-dependent variation of regions in epitaxy

The 2D process simulation is effected in arbitrarily polygonal-bounded and single-connected regions. The discretization of the boundaries by polygons corresponds to a given accuracy (in general in a not-equidistant way), which enables an effecient description of corplex regions.
In epitaxy the time-dependent variation of the region is done by a string algorithm. Discretization of the boundary is refined and a velocity is computed for each discretization point. For epitaxy, the direction of velocity is given by the outward normal and the value of velocity depends on the process parameters. In consideration of a time step-size a displacement vector is given in each discretization point and thus provisional new boundaries. This is followed by an algorithm for loop elimination. Also the discretization of the new boundaries is adapted to the given accuracy (s. picture 1).

3. Diffusion equation for impurities in epitaxy

For the technological step of epitaxy the diffusion equation has to be solved as a parabolic differential equation. Here only the boundary condition of the silicon/gas surface in epitaxy are considered (s. figure 2).
The main gas stream of an epitaxial reactor transports silicon and dopant compounds. Between main gas stream and the wafer there is a stagnant gas layer in which the above compounds diffuse with the flux

$$\vec{F}_Z = k_m \cdot (P_D^0 - P_D^*) \cdot \vec{e}_x \ .$$

Before inserted in the silicon lattice, the dopants are stored in the adsorbed layer on the silicon surface. For the stream F_S of impurities from the gas layer built in the adsorbed layer a stationary flow is assumed, namely

$$\vec{F}_S(r) \cdot \vec{n} = F_S(r) = k_f \cdot (P_D^* - C_{OF}(r)/k_p) \ .$$

The mass balance for the stagnant gas layer has obtained a steady state; this yields

$$\gamma_{all} F_Z = \int_{OF} F_S(r)\, dr \ .$$

In addition to /1/ a locally varied insertion flux into the adsorbed layer is assumed.
Now, these equations can be used to obtain

$$F_S(r_0) = k_{mf} \cdot \left[\frac{C_{OF}(r_0)}{k_p} - P_D^0 + \frac{k_f}{k_m k_p \gamma_{all}} \cdot \int_{OF} C_{OF}(r_0) - C_{OF}(r)\, db \right] ,$$

with

$$k_{mf} = \frac{\gamma_{all}/b \cdot k_m \cdot k_f}{\gamma_{all}/b \cdot k_m + k_f} \ .$$

This is the new boundary condition, which is used in the algorithm.

4. Algorithms for the solution of the impurity diffusion problem

For the simulation of epitaxy, the algorithm in figure 3 is used. The variation of the region is done according to paragraph 2. In the next 3 steps of this algorithm (as well as in other process steps in DUPSIM /2/), for space discretization the cell method and for time discretization the implicit Euler method are used.
For epitaxy, the time-dependent region has to be taken into consideration. The concentration of dopants from one time step to another is computed in assuming the condition that no dopands are lost. The formula is

$$c_i^{n+1} = \frac{\sum_{j=1}^{N^n} c_j^n * (V_j^n \cap V_i^{n+1})}{V_i^{n+1}} \ ,$$

where C_k is the concentration, V_k the volume of the cell, N the number of all discretization points and n and (n+1) are two time-steps. For discretization points according to silicon surface, it is corrected by

$$C_2^{n+1} := C_1^{n+1} V_1^n \cdot (1+V_{1ad}/V_1^n) / (V_1^{n+1}+V_{1ad})$$

because of storage of dopands in the absorption layer. In comparison to diffusion in time-dependent regions, the volume increased by V_{iad} is used for the cell V_i.

5. Example for epitaxy simulation

The application of the epitaxy model will be shown by means of the example of a low resistance-buried region, with the simple silicon wafer being the basis. It is structured by lithography in such a way that the silicon is etched and antimony is implanted on the right hand side (figure 4). Then, all layers on silicon are removed and an inert diffusion is effected.
The silicon will be deposited in the epitaxy step within time 5.0min, of a velocity 0.4µm/min and a temperature 1150°C. Phosphorus from the gas stream is built in. Figure 5 shows the result of this step. The antimony profile on the left-hand side is caused by lateral autodoping. The vertical autodoping results in a greater slope of profile in the epitaxy layer than in the substrat.

Literature

/1/ R.Reif, R.W.Dutton: "Computer Simulation in Silicon Epitaxy"; J. of Electrochem. soc., vol. 128 (1981) No.4, pp: 909-918

/2/ Elschner, H., Bothmann, D., Klix, W., Spallek, R.G., Stenzel, R., Vanselow, R., Voigt, R.: "Methods and Results in 2d Process and Device Simulation as well as in 3d Device Simulation"; NASECODE VI (1989), pp. 148-153

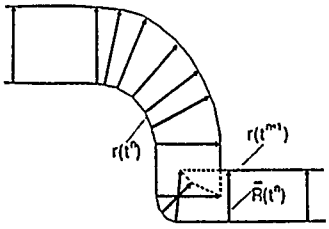


Figure 1: string algorithm to vary the structure

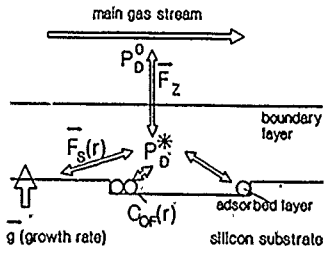


Figure 2: transports of dopants in epitaxy

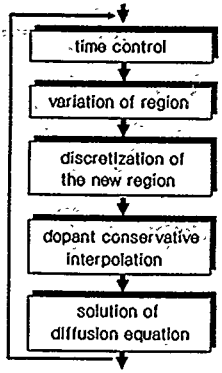


Figure 3: algorithm for epitaxy simulation

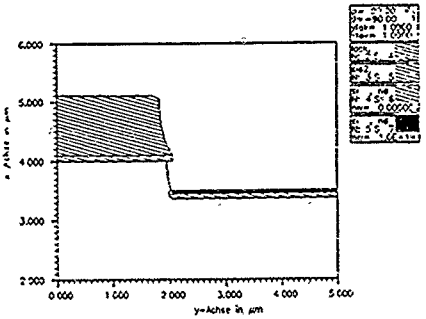


Figure 4: structure after implantation

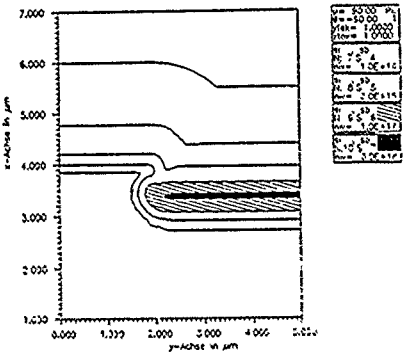


Figure 5: structure after epitaxy

Bounds for the solution of a system of parabolic equations arising in circuit theory

PEKKA NEITTAANMÄKI

Depart. of Mathematics
University of Jyväskylä
Seminaarinkatu 15
40100 Jyväskylä, Finland

CORNELIU A. MARINOV*

Depart. of Mathematics
University of Jyväskylä
Seminaarinkatu 15
40100 Jyväskylä, Finland

Abstract. When one studies the influence of interconnections on the delay time of MOS-integrated circuits, a system of parabolic equations coupled by boundary conditions appears. Lower and upper bounds for the solution are inferred and verified by numerical integration.

1. INTRODUCTION

Let us consider a "mixed type" circuit (see Fig. 1).

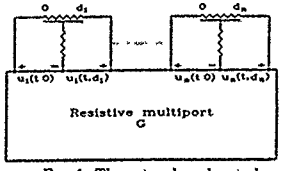


Fig. 1. The network under study

It consists of n *rcg*-transmission lines connected to a $2n$-resistive port. The lines are described by one-dimensional telegraph equations and the multiport is modelled by the matrix G. The sources are supposed to be switched off at the moment $t = 0$. These lead to the following mathematical model, (see [3-5]):

- a system of linear parabolic equations (u_i is the voltage along the i-th line):

$$\frac{\partial u_i}{\partial t} = a_i \frac{\partial^2 u_i}{\partial x^2} - b_i u_i\,, \qquad t \geq 0,\quad x \in (0,d_i)\,,\quad i = 1,2,\dots,n\,. \tag{1}$$

- a system of boundary conditions coupling the above equations:

$$\begin{bmatrix} h_1 \frac{\partial u_1}{\partial x}(t,0) \\ -h_1 \frac{\partial u_1}{\partial x}(t,d_1) \\ \vdots \\ h_n \frac{\partial u_n}{\partial x}(t,0) \\ -h_n \frac{\partial u_n}{\partial x}(t,d_n) \end{bmatrix} = G \begin{bmatrix} u_1(0) \\ u_1(d_1) \\ \vdots \\ u_n(0) \\ u_n(d_n) \end{bmatrix} \tag{2}$$

the initial conditions:

$$u_i(0,x) = u_{i_0}(x), \quad x \in [0,d_i],\ i = 1,2,\dots,n \tag{3}$$

This kind of problem appears when we study the influence of interconnection wires to the performances of a MOS integrated chip, especially from "the delay-time" point of view [1,2] Till now we have obtained results concerning the existence and uniqueness of this problem [3,4,7], as well as evaluations of the decay of the solution [5,6] even for certain nonlinear cases.

Here we are interested again in asymptotic behaviour predicting, giving upper bounds of each component of the solution (not a global bound as till now) but also lower bounds of exponential type.

* On leave from Faculty of Electrotechnics,
Polytechnical Institute of Bucharest, 77206 Bucharest, Romania

2. RESULTS

Let us denote for each $i = 1,2,\dots,n$, by α_i the solution in $[0,\pi/2)$ of the equation

$$\alpha_i \tan\alpha_i = \frac{d_i}{2h_i} \max\{(G_{2i-1,2i-1} + G_{2i-1,2i});(G_{2i,2i} + G_{2i,2i-1})\}$$

We set

$$\phi_i(x) = \cos\frac{d_i - 2x}{d_i}\alpha_i \text{ for } x \in [0,d_i] \text{ and } \lambda_i = 4\frac{a_i\alpha_i^2}{d_i^2} + b_i\,.$$

Also, for every $i = 1,2,\dots,n$, denote by β_i the solution in $[0,\pi/2)$ of the equation $\beta_i \tan\beta_i = \frac{d_i}{2h_i}\min(\sum_{k=1}^{2n} G_{2i-1,k};\ \sum_{k=1}^{2n} G_{2i,k})$ and let $\beta = \min\{\beta_i,\ i = 1,2,\dots,n\}$ and $\gamma = \min\{4\frac{\beta^2 a_i}{d_i^2} + b_i,\ i = 1,2,\dots,n\}$. Also, $\Psi_i(x) = \cos\frac{d_i - 2x}{d_i}\beta$. Now we are ready to formulate the result.

THEOREM 1.
Let us assume that the following hold.

1. $a_i > 0,\ b_i \geq 0,\ h_i > 0$ *for all* $i = 1,2,\dots,n$,
2. $G_{ij} \leq 0$ *for all* $i,j = 1,2,\dots,2n,\ i \neq j$,
$$\sum_{j=1}^{2n} G_{ij} \geq 0 \text{ for all } i = 1,2,\dots,2n\,,$$
3. $u_{i0}(x) \geq 0$ *for* $x \in [0,d_i]$ *and* $i = 1,2,\dots,n$.

Then, if $u = (u_1,\dots,u_n)$ *is the solution of (1)+(2)+(3) with* $u_i \in C([0,T] \times [0,d_i])$ *and* $\frac{\partial u_i}{\partial t},\ \frac{\partial^2 u_i}{\partial x^2} \in C((0,T) \times (0,d_i))$, *we have for each* i, *and* $x \in [0,d_i]$:

$$u_i(t,x) \geq e^{-\lambda_i t}\phi_i(x) \min_{0\leq x\leq d_i} \frac{u_{i0}(x)}{\phi_i(x)}$$

and

$$u_i(t,x) \leq e^{-\gamma t}\Psi_i(x) \max_{\substack{0\leq x\leq d_i \\ 1\leq i\leq n}} \frac{u_{i0}(x)}{\Psi_i(x)}$$

3. SKETCH OF PROOF

The proof is based on the maximum principle for parabolic operators, the first step being to prove that $u_i(t,x) \geq 0$ for all i, t, x Let us denote $m = \inf\{u_i(t,x);\ t \in [0,T],\ x \in [0,d_i],\ i = 1,2,\dots,n\} = u_p(t_m,x_m)$ and suppose $m < 0$ If $x_m = 0$, from odd rows of (2) we easily obtain $h_p \frac{\partial u_p}{\partial t}(t_m,0) \leq 0$ contradicting a known result [8, Ch 3,Th.4]. The same happens if we suppose $x_m = d_p$. Then the maximum principle give $m = u_p(0,x_m)$ from where we obtain the desired positivity.

To obtain the positive lower bound we change the functions $u_i(t,x) = v_i(t,x)e^{-\lambda_i t}\phi_i(x)$, and for the system with v_i as unknowns apply again the above reasoning, finding $v_i(t,x) \geq \min_{0\leq x\leq d_i} v_i(0,x)$ from where we have the lower bound. With the change $u_i(t,x) = v_i(t,x)e^{-\gamma t}\Psi(x)$ we obtain by a similar method $v_i(t,x) \leq \max_{\substack{0\leq x\leq d_i \\ 1\leq i\leq n}} v_i(0,x)$ from where we derive the upper bound

4. Numerical example

In order to verify how far are these bounds to the exact solution we have computed numerically many typical examples from MOS interconnections. We have used a finite element discretization in space combined to a ready-made subroutine from NAG library (residual evaluations method) for time integration.

Let us consider the tree-type network from Fig. 2

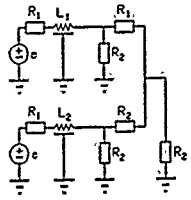


Figure 2. An example

The values of parameters are: $a_1 = 1/10$, $b_1 = 1/100$, $a_2 = 1$, $b_2 = 1/10$, $h_1 = h_2 = 1$, $d_1 = d_2 = 1$, for the lines and $R_1 = 1$, $R_2 = 10$ for the lumped part. From here we derive the G matrix: $G_{11} = G_{33} = 1$, $G_{22} = 32/120$, $G_{44} = 23/120$, $G_{21} = G_{42} = -1/12$, the remaining elements being zero.

The numerically computed components of the solution are shown in Figs. 3 and 4.

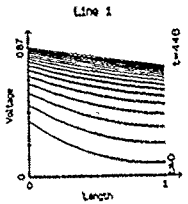


Figure 3 $u_1(t,x)$

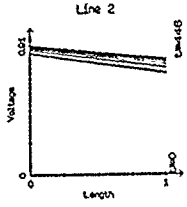


Figure 4. $u_2(t,x)$

The Figs. 5 and 6 present the solutions at the right hand end of the lines together with their bounds.

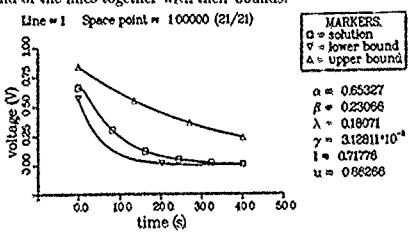


Figure 5 $u_1(t,d_1)$

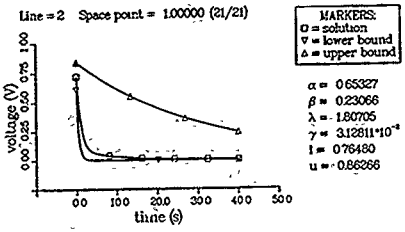


Figure 6 $u_1(t,d_2)$

5. Concluding remarks

If in assumption 2. we take $\sum_{j=1}^{2n} G_{ij} > 0$ for all $i = 1,2,\ldots,2n$, then all the components are globally exponential asymptotically stable even for the case $b_i = 0$. If $\sum_{j=1}^{2n} G_{ij} = 0$ for at least one $i = 1,2,\ldots,2n$, and for the same i we have $b_i = 0$, then $\gamma = 0$ and we obtain only constant upper bounds of all components. From this point of view, the lower bounds seem to be better. Indeed, if for a certain i we have $\max[(G_{2i-1,2i-1} + G_{2i-1,2i}),(G_{2i,2i} + G_{2i,2i-1})] = 0$ and $b_i = 0$ then we obtain a time-constant lower bound for u_i, while the other components may have exponential lower bounds. The most numerical examples seem to show also that the lower bound is tighter, but other examples have given an opposite result. Anyway, the bounds are sufficiently tight to be usefull for designers if we take into account their "global" feature and their easy computation.

References

[1] J.Rubinstein, P.Penfield, Jr. and M.A. Horowitz, *Signal delay in RC tree networks*, IEEE Trans Computer-Aided Design,July 1983 CAD-2, 202-211.

[2] J.L. Wyatt, Jr., *Signal delay in RC mesh networks*, IEEE Trans. Circ.Syst., May 1985 CAS-32, 507-510.

[3] C.A. Marinov and P. Neittaanmäki, *A theory of electrical circuits with resistively coupled distributed structures; Delay time predicting*, IEEE Trans. Circ. Syst., February 1988 CAS-35, 173-183.

[4] C.A. Marinov and A. Lehtonen, *Mixed-type circuits with distributed and lumped parameters*, IEEE Trans. Circ. Syst. CAS-36 (Aug 1989), 1080-1086

[5] C.A. Marinov and P. Neittaanmäki, *Global delay time for general distributed networks with applications to timing analysis of digital MOS integrated circuits*, Int J Comp Math. Electrical Electronic Eng.-COMPEL.8, Jan. 1989, 17-37.

[6] C.A. Marinov and P. Neittaanmäki, *Asymptotical convergence evaluation for a parabolic problem arising in circuit theory*, Zeitschrift für Angewandte Mathematik und Mechanik-ZAMM 70 (Aug. 1990), 344-347.

[7] G. Morosanu, C.A. Marinov and P. Neittaanmäki, *Well-posed nonlinear problems in integrated circuit modelling*, to appear in Circ Syst Signal Proc.

[8] M.H. Protter and H.F. Weinberger, "Maximum principles in differential equations," Springer-Verlag, N.Y., 1984.

A fixed domain approach in an optimal shape design problem

D. TIBA
Inst. of Math., Romanian Academy
of Sciences, Bucuresti 79622, ROMANIA
and INRIA 78153, Le Chesnay Cedex
FRANCE

P. NEITTAANMÄKI and R. MÄKINEN*
Department of Mathematics
University of Jyväskylä
PL 35, SF-40351 Jyväskylä
FINLAND

Abstract. A fixed domain approach is presented for solving optimal shape design problems. In the proposed method the original optimal shape design problem is converted to a control problem settled in a fixed domain. The method is demonstrated in solving an optimal shape design problem arising from transmission problems. Results of numerical tests are presented.

Keywords. Optimal shape design, control approach.
Subject classification: 49E99, 35J67

1. INTRODUCTION

The standard way to solve optimal shape design problems numerically is the boundary variation method. In that method the unknown boundary is parametrized using a set of design parameters [4], [5].

Here, we discuss another approach suggested by recent controllability-type results for elliptic systems [1], [6], [7]. It may be mainly compared with the method of mapping or with the fictitious domain method [3]. In the method of mapping the problem is converted to a control problem in a fixed domain and control in coefficients whereas in fictitious domain method the control is on the right hand side.

We shall convert the problem to a control problem in a fixed domain, with control in the coefficient. However, unlike in the method of mapping or fictitious domain method, the topology of the variable domains is not given a priori.

Our basic idea is a simple one: if Ω is a regular subdomain of a fixed domain D, then it is possible to find some mapping $p : D \to \mathbf{R}$ (by an exact controllability-type argument [7]) such that

$p > 0$ in Ω, $p = 0$ on $\partial\Omega$ and $p < 0$ in $D \setminus \overline{\Omega}$. Then the Heaviside mapping $H : \mathbf{R} \to \mathbf{R}$

$$H(p) = \begin{cases} 1, & p \geq 0, \\ 0, & p < 0, \end{cases} \tag{1.1}$$

is the characteristic function of $\overline{\Omega}$ in D.

In the sequel, we shall apply this approach to a model problem discussed by Céa [2] and Pironneau [5, Ch. 8]. In section 2, we perform a brief theoretical analysis of the proposed method and some numerical results are given in section 3.

2. THE MAIN RESULTS

We study the following optimization problem

$$\underset{\Omega \subset D}{\text{minimize}} \int_E |y_\Omega - y_d|^2 \, dx \tag{P}$$

subject to the transmission problem

$$\begin{aligned} -a_1 \Delta y_1 + a_0 y_1 &= f \text{ in } \Omega, \\ -a_2 \Delta y_2 + a_0 y_2 &= f \text{ in } D \setminus \overline{\Omega}, \\ a_1 \frac{\partial y_1}{\partial n} = a_2 \frac{\partial y_2}{\partial n} &\quad \text{in } \partial\Omega \setminus (\partial\Omega \cap \partial D), \\ y_1 = y_2 &\quad \text{in } \partial\Omega \setminus (\partial\Omega \cap \partial D), \\ a_i \frac{\partial y_i}{\partial n} = 0 &\quad \text{in } \partial D, \ i = 1, 2 . \end{aligned} \tag{T}$$

Above a_0, a_1, a_2 are positive constants, $\frac{\partial}{\partial n}$ denotes the exterior normal derivative to Ω or D, $y_d \in L^2(E)$, $f \in L^2(\Omega)$,

$E \subset D$ is a fixed measurable subset and $y_\Omega \in H^1(\Omega)$ is given by

$$y_\Omega(x) = \begin{cases} y_1(x) & \text{in } \Omega, \\ y_2(x) & \text{in } D \setminus \Omega . \end{cases} \tag{2.1}$$

*Research supported by the Academy of Finland.

If χ is the characteristic function of Ω in D, then the variational formulation of the problem (T) is given by

$$\int_D ([a_1\chi + a_2(1-\chi)]\nabla y_\Omega \cdot \nabla w + a_0 y_\Omega w - fw)\, dx = 0 \quad \forall\, w \in H^1(D) \tag{2.2}$$

and, in [5], it is analysed the case when

$$\chi \in \{g : D \to \mathbf{R} \mid g(x) = 0 \text{ or } g(x) = 1 \,\forall\, x \in D\} \tag{2.3}$$

is the control parameter. As the form of the constraint (2.3) makes the problem difficult to handle we replace (2.2) by

$$\int_D ([a_1 H(p) + a_2(1 - H(p))]\nabla y_\Omega \nabla w + a_0 y_\Omega w - fw)\, dx = 0, \quad \forall\, w \in H^1(D). \tag{2.4}$$

We approximate (2.4) by

$$\int_D ([a_1 H_\varepsilon(p) + a_2(1 - H_\varepsilon(p))]\nabla y \nabla w + a_0 y w - fw)\, dx = 0, \quad \forall w \in H^1(D), \tag{2.5}$$

where H_ε is the Yosida approximation of the maximal monotone extension of H in $\mathbf{R} \times \mathbf{R}$.

THEOREM 2.2. *Let $y_\varepsilon \in H^1(D)$ be the unique solution of (2.5). Then $y_\varepsilon \to y_\Omega$ strongly in $H^1(D)$, when $\varepsilon \to 0$.*

We approximate the problem (P), (T) by the following one:

$$\text{minimize} \int_E |y - y_d|^2\, dx, \qquad (\mathbf{P}_\varepsilon)$$

subject to any measurable p and $y \in H^1(D)$ given by (2.5).

REMARK 2.3. Generally, in the absence of some compactness assumption on the class of subdomains Ω (for instance ε-cone property, Pironneau [5, Ch.3]), one may not obtain the existence of a solution for the problem (P), (T). The same is valid for the problem $(\mathbf{P}_\varepsilon)$ since there are no coercivity conditions on the control parameter p. Obviously, one may ask a boundedness condition on p, $|p(x)| \le 1$, due to the relationship between p and Ω. But this does not imply existence since the weak limit in $L^\infty(D)$ of a sequence of characteristic functions is not necessarily a characteristic function.

We denote $[y^\varepsilon, p_\varepsilon]$ an δ-optimal pair of the problem $(\mathbf{P}_\varepsilon)$, $\delta > 0$, that is:

$$J_\varepsilon(y^\varepsilon, p_\varepsilon) \le \inf(\mathbf{P}_\varepsilon) + \delta, \tag{2.6}$$

where

$$J_\varepsilon(y^\varepsilon, p_\varepsilon) = \int_E |y^\varepsilon - y_d|^2\, dx$$

and

$$\int_D ([a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla y^\varepsilon \nabla w + a_0 y^\varepsilon w - fw)\, dx = 0, \quad \forall w \in H^1(D). \tag{2.7}$$

PROPOSITION 2.4. *For $\varepsilon \to 0$, we have $y^\varepsilon \to \tilde{y}$ strongly in $L^2(D)$ and weakly in $H^1(D)$ on a subsequence, such that*

$$\int_E |\tilde{y} - y_d|^2\, dx \le \inf(\mathbf{P}) + \delta \tag{2.8}$$

In order to solve the problem $(\mathbf{P}_\varepsilon)$ by a gradient-type method, we discuss the adjoint system. Let $\theta_\varepsilon : L^\infty(D) \to L^2(D)$ be the approximate state mapping $p \to y$ defined by (2.5) (we restrict p to be in $L^\infty(D)$ here).

PROPOSITION 2.6. *θ_ε is a Gateaux differentiable mapping and $\nabla\theta_\varepsilon(p)v = r$ satisfies*

$$\int_D \big[(a_1 - a_2)H'_\varepsilon(p)v\nabla y \cdot \nabla w + (a_1 H_\varepsilon(p) + a_2(1 - H_\varepsilon(p)))\nabla r \cdot \nabla w + a_0 r w\big]\, dx = 0 \quad \forall\, w \in H^1(D), \tag{2.9}$$

where $y = \theta_\varepsilon(p) \in H^1(D)$ and v is arbitrary fixed in $L^\infty(D)$. Moreover $r \in H^1(D)$.

3. NUMERICAL EXAMPLES

In this section we choose $D = E = (0,1) \times (0,1)$, $a_1 = 10$, $a_2 = a_0 = 1$ and $f = x_1^2 + x_2^2$. The Heaviside mapping is approximated by

$$H_\varepsilon(p) = \begin{cases} 1 - \frac{1}{2}e^{-p/\varepsilon}, & p \ge 0, \\ \frac{1}{2}e^{p/\varepsilon}, & p < 0. \end{cases} \tag{3.1}$$

The regularized state problem (2.5) is discretized by the finite element method. The control parameter p in (P_ε) is taken to be piecewise constant.

In the numerical solution we have used a gradient algorithm from the NAG-subroutine Library. In computations the cost function was scaled with a factor 0.5×10^5 and the regularization parameter $\varepsilon = 1/10$ was used. As the discrete control parameter p is piecewise constant, nodal averaging of it was done for plotting purposes.

Example 3.1. Let y_d be the solution of the transmission problem in the geometry shown in Figure 3.1. Two runs with 100 and 900 finite elements were done. In both cases initial guess $p = 0$ was used. The results are shown in Table 3.1 and Figures 3.2–3.3 (contour $p = 0$). In both cases the global optimums were clearly found.

Elem.	In. cost	Final cost	Iter.	CPU-s.
100	23.1	1.41×10^{-3}	15	29
900	24.9	1.11×10^{-3}	23	395

Table 3.1

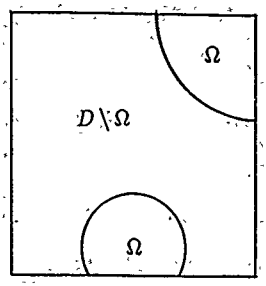


Figure 3.1

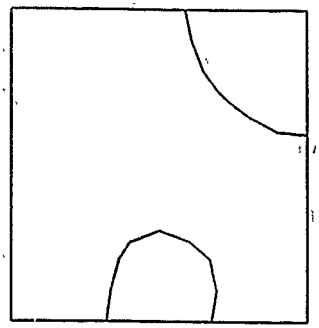


Figure 3.2

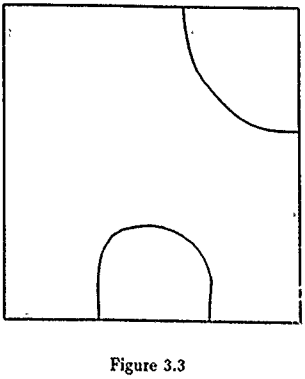


Figure 3.3

References

[1] V. Barbu and D. Tiba, *Boundary controllability for the coincidence set in the obstacle problem*, to appear in SIAM J. Control and Optimiz. (1991).

[2] J. Céa, *Identification de domaines*, in "Proc. of 5th Conference on Optimization Techniques, Part I" (G. Goos and J. Hartmanis, eds.), Lecture Notes in Computer Science 3, Springer Verlag, Berlin (1973), 92-102.

[3] J. Haslinger, K.-H. Hoffmann and M. Kočvara, *Control/Fictitious domain method for solving optimal shape design problems*, (to appear).

[4] J. Haslinger and P. Neittaanmäki, "Finite element approximation of optimal shape design. Theory and Applications," J. Wiley& Sons, Chichester, 1988.

[5] O. Pironneau, "Optimal shape design for elliptic systems," Springer-Verlag, Berlin, 1984.

[6] D. Tiba, *Une approche par controlabilité frontière dans les problémes de design optimal*, C.R.A.S. Paris t. 310 Séries I (1990).

[7] D. Tiba, P. Neittaanmäki and R. Mäkinen, *Controllability type properties for elliptic systems and applications*, to appear in "Optimal Control of Distributed Systems", F. Kappel, K. Kunisch (eds.), Birkäuser Verlag Basel (1991).

An Accurate and Efficient Delay Time Modelling and its Application to CMOS Data Path Evaluation and Transistor Sizing

D. AUVERGNE, N. AZEMARD, D. DESCHACHT, M.ROBERT

Université de Montpellier II : Laboratoire d'Automatique et de Microélectronique de Montpellier (UA D03710 CNRS)
Pl. E. Bataillon, 34095 MONTPELLIER Cedex 5, FRANCE (Telex: 490944F)

Abstract– In this paper we show how a simplified model of CMOS structures switching conditions can be used to obtain explicit delay formulations, with clear evidence of relevant parameters. Applications to a data path evaluation and to the definition of a local strategy for transistor sizing are given as necessary facilities for a cell synthetizer.

I. Introduction

If top down techniques allow to control the complexity found in digital circuit designs, physical issues, such as layout style, transistor size, wiring and timing are too much important for the quality of the resulting design to be ignored until the last implementation step. As part of the design automation project conducted in the Microelectronics Department (LAMM) of Montpellier, we are developing a technology independent cell synthetizer with automatic performance evaluation facilities [1].

Efficient structural synthesis demands the use of prior structural criteria as an effective initial solution to layout implementation and to evaluate propagation delays across data paths, whilst taking into account device sizes, process parameters, layout parasitics, input waveform from driving cells and active and wiring output loading. In order to do this, simplified delay models with clear evidence of structural, technological and environmental parameters must be defined at the transistor level.

The aim of this paper is to present an accurate analytical model for the propagation delay of CMOS gates and its use in the cell synthetizer for speed performance evaluation and transistor sizing.

II. Delay Modelling

Great efforts have been made to develop a fast and accurate timing analysis [2,5]. Simplified RC models [6] have, until now, been limited to unidirectional structures with a low degree of accuracy in treating transistor serial arrays. Common alternative has been to use delay table representation [3] and polynomial decomposition [7] which, as technology dependent solutions, are limited to standard unit evaluation allowing the propagation of worst or best case delays. Moreover, without explicit evidence of the physical parameters, these models can only be used to develop tools for the verification of delay specifications and do not allow the definition of sizing and optimization criteria.

By dividing the data path into unidirectional element (including transmission gates associated to their power source), it appeared possible [8,9], from the physical modeling of the individual transistor's switching operations, to obtain a real delay evaluation of ANDORI from a linear combination of the driving (i-1) and the controlled (i) structure's step responses, as follows:

$$t_{HL}(i) = \frac{A \cdot t_{HLs}(i-1) + t_{HLs}(i)}{1 + \alpha \cdot A \cdot \frac{t_{HLs}(i-1)}{t_{HLs}(i)}} \qquad (1)$$

$$t_{LH}(i) = \frac{B \cdot t_{HLs}(i-1) + t_{LHs}(i)}{1 + \alpha \cdot B \cdot \frac{t_{HLs}(i-1)}{t_{LHs}(i)}}$$

where :
- A, B are linearization coefficients,
- α is an input slope correcting factor [10],
- t_{HLS}, t_{LHS} are the fall and rise step responses of general ANDORI, evaluated for an evolution of the output voltage from the static level to Vcc/2 [9]

These step responses can be obtained directly from the mean charge transfer, evaluated across the node under

consideration and produced by the imbalance current developed in the cell under evaluation, as follows:

$$t_{HL}(i) = \tau_{ST} \cdot \frac{C_L}{2\,C_N(i)}$$

$$t_{LH}(i) = \tau_{ST} \cdot \frac{\mu_N}{\mu_P} \cdot \frac{C_L}{2\,C_P(i)} \quad (2)$$

for an inverter made up of N (C_N) and P(C_P) transistors loaded by a capacitance C_L, where :

$$\tau_{ST} = \frac{2Cox\,W_N\,Lmin^2}{\mu_N\,Cox\,W_N\,(Vcc-Vt)} \cdot \frac{8Vcc\,(Vcc-Vt)}{7Vcc^2 + 4Vt^2 - 12VccVt} \quad (3)$$

is the elementary fall time characteristic of the technology and :

$$t_{HL,LH} = tinv_{HL,LH} + \frac{Reff\,Co}{nCo+C_L}\left(\frac{Co}{12}n\,(2n^2+1) + \frac{n^2C_L}{2}\right) \quad (4)$$

represents the step response of an inverter loaded by a serial array of transistors (TG).

This expression has been obtained while considering that during the switching process the TG are working in linear mode and can be modeled by a resistance, R_{eff} and a capacitance Co. As given in equ.2, t_{INV} represents the step responses of the driving inverter loaded by the complete array equivalent load, C_L+nCo. The second term of equ. 4 represents the propagation delay on the array, of the charge variation imposed by the sourcing device.

Equ.1 and 4 can easily be used to express the real responses of general ANDORI as the product of 3 terms :

– a temporal scaling factor τ_{st}, characteristic of the technology,

– a structural factor (C_N, C_P, n)characteristic of the switching network,

– an environmental factor including the loading factor (generalized fanout term) specifying the total active and parasitic (structural and interconnection) load, and the controlling device.

III. Application to CMOS Data Path Evaluation

The use of these expressions to evaluate delays across various data paths has resulted in values which are very close to those obtained by electrical simulations (discrepancies of less than 10 %) [9,7,10]. These equations have been integrated into a timing predictor (PATH RUNNER [11]), allowing a fast and accurate post layout evaluation of the speed performances of synthetized cells. Figure 1a and 1b summarize the speed and accuracy performances obtained by characterizing a 2µm CMOS library.

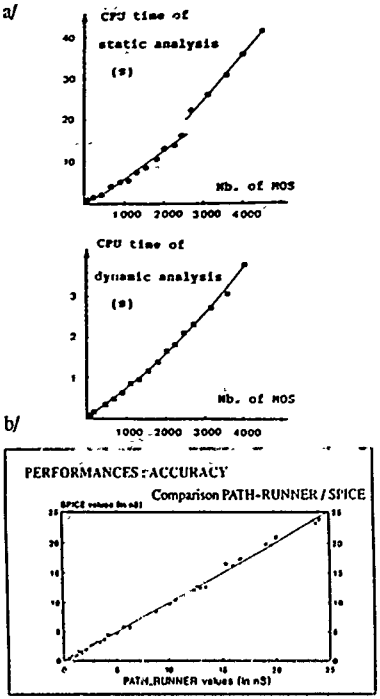


Figure 1 : Illustration of the characteristics of the PATH-RUNNER timing evaluator :
a) speed performances are evaluated on SUN 3/60 (1.5 MIPS),
b) delay values evaluated on different data paths are compared to values obtained from SPICE simulations (2µm).

IV. Application to Transistor Sizing

Written in reduced form with respect to a capacitance and used as a reference, equation 4 gives direct information about the relative delay cost of structural and loading choices, thus allowing the definition of explicit sizing laws.

For example, the sizing of an irregular inverter array is directly obtained from the solution of :

$$\left(1+\frac{\mu_N}{\mu_P}\right)\frac{X_N(i)}{X_N(i-1)}=\frac{Y_i+\left(1+\frac{\mu_N}{\mu_P}\right)\cdot X_N(i+1)}{X_N(i)} \qquad (5)$$

where X_N, Yi are the values of N transistor active capacitance and parasitic load normalized with respect to the reference capacitance.

This represents a system of dependent equations ($X_N(i)^2 \approx X_N(i+1) \cdot X_N(i-1)$) with a complexity equal to the number of elements to be sized. A global solution can be obtained through successive iterations of the initial solution, but major difficulties are met in solving divergence branches.

As an effective speed area trade off, we have proposed a local solution allowing the backward processing of the array from the output load to the input. This local solution is defined, sizing each cell with respect to a reference cell, which will define the area delay trade off. For example, using a minimum size inverter as a reference (N transistor reduced width : $X_N(0)=1$), it can be shown [12] that the local sizing law, deduced from equ.5 :

$$X_N(i) = \frac{1}{\sqrt{1+\frac{\mu_N}{\mu_P}}}\cdot\sqrt{Y_i \cdot X_N(0)} \qquad (6)$$

allows the sizing of an array of inverters with a smaller area (50% less), at the expense of 10% extra delay, with respect to the global sizing that we can obtain using a mathematical minimization algorithm [13].

These results can easily be generalized to ANDORI, allowing a fast initial estimation of transistor sizes corresponding to a minimum area implementation with nearly optimal delay.

V. Conclusion and Future Developments

We have presented a physical modelization of delays on CMOS structures, allowing the fast and accurate evaluation of delays and the definition of efficient sizing laws. These results have been applied to a cell synthetizer for automatic transistor sizing and evaluation of performances. The parametrization of a layout induced parasitic associated to structural choices is under development as an efficient way to drive structural synthesis.

References

[1] M. ROBERT, D. DESCHACHT, G. CATHEBRAS, S. PRAVOSSOUDOVITCH, D. AUVERGNE : *"PRINT methodology a compilation approach for cell library generation"*, 1988 IEEE Symp. on Circuits and Systems Espoo, FINLAND, oct. 1988.

[2] N.P JOUPPI : *"Timing analysis and performance improvements of MOS VLSI designs"*, IEEE trans. on CAD, vol CAD 6, n° 4, 1987.

[3] J.K. OUSTERHOUT : *"CRYSTAL : a timing analyser for NMOS VLSI circuits"*, Proc. 3rd Caltech VLSI conf R.BRYANT, ed. 1983.

[4] C.F CHEN, C. LO, H.N. NHAM, P. SUBRAMANIAM *"The second generation MOTIS mixed mode simulation"* 21st Design Automation Conf., Albuquerque, 1984.

[5] M.R. DAGENAIS, C. RUMIN : *"Timing analysis and verification of digital MOS circuits"* , Conf. on VLSI and computers, Hamburg, may 1987.

[6] P. PENFIELD, J. RUBINSTEIN, M.A. HOROWITZ *"Signal delay in RC tree networks"*, IEEE transactions on CAD, vol. CAD 2, n° 3, 1983.

[7] Y.H. JUN, K. JUN, S.B. PARK: *"An accurate efficient delay simulator for MOS logic circuits using polynomial approximation"*, IEEE proc.Int. Symp. on Circuits and Systems, Vol. 3, FINLAND, 1988.

[8] D. AUVERGNE, D. DESCHACHT, M. ROBERT. *"Explicit formulation of delays in CMOS VLSI"*, Electronics letters, vol. 23, n° 14, 1987.

[9] D. DESCHACHT, M. ROBERT, D. AUVERGNE *"Explicit formulation of delays on CMOS data path"* IEEE J. of Solid state circuits, vol. 23, n° 5, oct. 1988.

[10] D. AUVERGNE, D. DESCHACHT, M ROBERT *"Input waveform slope effects in CMOS delays"*, IEEE Solid State circuits, vol. 25, n° 6, dec. 1990.

[11] D. DESCHACHT, P. PINEDE, M. ROBERT, D. AUVERGNE: "PATH RUNNER : *an accurate and fast timing analyser"*, EDAC, Glasgow, march 1990.

[12] D. AUVERGNE, N. AZEMARD, V. BONZOM, D DESCHACHT, M. ROBERT: *" Formal sizing rules of CMOS circuits"*, EDAC, Amsterdam, february 1991.

[13] D.M.HIMMELBLAU, Applied non linear programming, *"Rosenbrock's method"* chp. 4, Mc Graw Hill ed. 1972.

Constrained Approximation of Dominant Time Constant(s) in RC Circuit Delay Models *

Nanda Gopal, Curtis L. Ratzlaff, and Lawrence T. Pillage
Department of Electrical & Computer Engineering
The University of Texas at Austin
Austin, Texas 78712

1 Introduction

Linear RC circuits, particularly RC trees, have been very popular for modeling the signal propogation and delay in digital integrated circuits. A linear RC tree is a reasonable model of the physical interconnect of an integrated circuit. More importantly, the RC tree response waveforms are guaranteed monotone and therefore can be analyzed with extreme efficiency in terms of the first moment of the impulse response, also known as the Elmore delay [4] One can compute the Elmore delay for an RC tree with path tracing, in linear time. This efficiently provides a dominant time constant, or first-order waveform approximation for the RC tree response. Furthermore, best/worst case bounds can be obtained with equal efficiency [12]. Recently, with multilayer, submicron interconnects, more generalized RC circuits have been used to model on-chip signal propogation. The effects of dominant zeros and non-monotone waveshape behavior require the use of higher-order approximations [3, 9].

Asymptotic Waveform Evaluation (AWE) [9], an nth-order extension of the Elmore approximation for RC trees, is one such higher-order approximation for RLC circuits. AWE performs model-order reduction by a moment matching technique recognized to be a form of partial Pade' approximation [6]. To obtain a qth order, i.e., a q time constant, approximation, $2q$ moments drawn from the actual circuit and the model are matched. Moment-matching methods, however, are prone to yielding unstable models of stable systems [2]. To avoid unstable approximations, it is suggested that higher-order approximations be tried [2]. Unfortunately, due to finite machine precision, positive time constants are even more likely at higher orders. This paper proposes a constrained optimization for mapping moments to stable dominant time constants in AWE for RC interconnect models.

Section 2 briefly reviews the development of the equations in AWE that form the objective functions in this work. The proposed optimization scheme is described in section 3, followed by some results in section 4 and concluding remarks in section 5.

*This work was supported in part by the National Science Foundation MIP-9007917 and the Semiconductor Research Corporation under contract #90-DP-142.

2 Background

A brief description of AWE is provided below. For a complete discussion, the reader is referred to [8, 9].

The mathematical development of AWE is based on the differential state equations for a lumped, linear, time-invariant circuit:

$$\dot{\vec{x}} = \mathbf{A}\cdot\vec{x} + \mathbf{B}\cdot\vec{u}, \tag{1}$$

where $\vec{x}$ is the n-dimensional state vector and $\vec{u}$ is the m-dimensional excitation vector. For the investigation of delay and rise-time effects, a step excitation is adequate and assumed. For this excitation, (1) has the homogeneous solution

$$\dot{\vec{x}}_h = \mathbf{A}\cdot\vec{x}_h \tag{2}$$

with the initial condition

$$\vec{x}_h(0) = \vec{x}_0 + \mathbf{A}^{-1}\cdot\mathbf{B}\cdot\vec{u}_0 \tag{3}$$

where $\vec{x}_0$ is the initial state at time zero Taking the Laplace transform of (2) yields

$$\mathbf{X}_h(s) = (s\mathbf{I}-\mathbf{A})^{-1}\cdot\vec{x}_h(0). \tag{4}$$

The MacLaurin series expansion of (4) is

$$\mathbf{X}_h(s) = -\mathbf{A}^{-1}(\mathbf{I}+\mathbf{A}^{-1}s+\mathbf{A}^{-2}s^2\ldots)\vec{x}_h(0). \tag{5}$$

From (5), the initial conditions and first 2q-1 time moments of the i^{th} component of $\mathbf{X}_h(s)$ are characterized as

$$\begin{aligned}(\vec{m}_{-1})_i &= (-\vec{x}_h(0))_i\\ (\vec{m}_0)_i &= (-\mathbf{A}^{-1}\cdot\vec{x}_h(0))_i\\ &\vdots\\ (\vec{m}_{2q-2})_i &= (-\mathbf{A}^{-2q+1}\cdot\vec{x}_h(0))_i\end{aligned} \tag{6}$$

where the initial conditions are represented as the negative first moment. For RC interconnects, these moments $\vec{m}$ are computed efficiently using a path tracing technique described in [11], of which this work is a part

The reduced-order model has the form

$$\hat{X}_i(s) = -\sum_{l=1}^{q}\frac{k_l\eta}{1-s\eta} \tag{7}$$

where η_l and k_l represent the dominant time constants and corresponding residues respectively, and q represents the order of the reduced model. It now remains to solve for these time constants and residues by matching the moments from (6) to those of (7).

Expanding each of the terms in (7) into a series about the origin, and upon inclusion of initial conditions, the following set of nonlinear simultaneous equations for the i^{th} state variable is obtained.

$$\begin{aligned} -(k_1 + k_2 + \cdots + k_q) &= (\vec{m}_{-1})_i \\ -(k_1\tau_1 + k_2\tau_2 + \cdots + k_q\tau_q) &= (\vec{m}_0)_i \\ &\vdots \\ -(k_1\tau_1^{2q-1} + \cdots + k_q\tau_q^{2q-1}) &= (\vec{m}_{2q-2})_i \end{aligned} \quad (8)$$

These equations can be expressed in matrix form as

$$\vec{k} = -\mathcal{V}^{-1} \cdot \vec{m}_l \quad (9)$$

$$\mathcal{V} \cdot \Lambda^{-q} \cdot \mathcal{V}^{-1} \cdot \vec{m}_l = \vec{m}_h \quad (10)$$

where $\vec{m}_l$ represents the low-order moments (-1, 0, ..., q-2), $\vec{m}_h$ represents the high-order moments (q-1, q, ..., 2q-2), Λ^{-1} is a diagonal matrix of the time constants, and $\mathcal{V}$ is the well-known Vandermonde matrix.

Attempts to directly solve the nonlinear equations (9) and (10) to obtain the required time constants and residues as described in [8] or by using unconstrained methods such as Newton-Raphson iteration may lead to the problem mentioned earlier, i.e., unstable models of stable systems. In this work, a constrained, optimal solution is sought that restricts the time constants obtained to the left half plane.

3. Development

A two step optimization scheme is proposed to fit the moments in (7) to those in (6): (i) a constrained optimization to obtain the stable time constants that best satisfy (10), and (ii) a least-squares fit of the overdetermined system in (9) using the time constants obtained earlier. (This would take into account the presence of round-off errors in the moment values and find a best fit.)

As derived in AWE, (10) is a nonlinear system of the time constants η_l only, and can be solved independently of (9). To ensure that the reduced-order model (7) is stable for a given stable system, in this case, an RC tree, it would require that the time constants be negative. This constraint can be incorporated in a simple manner, without significant complication of the equations in (10), by a transformation of variables [1], that tranform the constrained system into an unconstrained one.

$$\eta_l = -e^{x_l}. \quad (11)$$

The resulting equations in x_l can now be optimized using more powerful unconstrained schemes.

In this work, a gradient descent scheme, based on the nonlinear Newton-Raphson iteration, is used to obtain the time constants. However, rather than obtain an exact solution, the gradient information is used to descend to an extremal point, in this case, a minima. This was encouraged by the availability of a good initial guess obtained by the direct mapping described in [8]. When the approach in [8] yields positive time constants, the remaining time constants have been seen to lie very close to the correct solution. A reasonably good initial estimate is thus obtained from the stable time constants and the negative of the errant time constant(s).

Gradient information in the form of the Jacobian matrix can be simply computed as illustrated below. Rewriting the system of equations to be minimized (10)

$$\vec{f}(\vec{x}) = \mathcal{V} \cdot \Lambda^{-q} \cdot \mathcal{V}^{-1} \cdot \vec{m}_l - \vec{m}_h \quad (12)$$

an expression for the Jacobian matrix is derived as.

$$\frac{\partial \vec{f}(\vec{x})}{\partial x_l} = \left(\frac{\partial(\mathcal{V} \cdot \Lambda^{-q})}{\partial x_l} - \mathcal{V} \cdot \Lambda^{-q} \cdot \mathcal{V}^{-1} \frac{\partial \mathcal{V}}{\partial x_l} \right) \vec{z} \quad (13)$$

where $\vec{z} = \mathcal{V}^{-1} \cdot \Lambda^{-q}$. This requires the evaluation of the partial derivatives of only the Vandermonde and the diagonal matrix, both of which yield sparse and extremely predictable derivatives.

With an optimal set of time constants in the left-half plane that minimize (12), it now remains to determine the corresponding residues in the model (7). While (9) provides a simple and straightforward solution, it uses only the lower-order moments. To further decrease the errors incurred in constraining the time constants, the overdetermined system (8) that matches all the $2q$ moments is used to obtain a best fit estimate using a least-squares approach.

Rewriting (8) in matrix form yields

$$\mathbf{L} \cdot \vec{k} = \vec{m} \quad (14)$$

where $\mathbf{L}$ is the $(q \times n)$ matrix

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \tau_1^1 & \tau_2^1 & \cdots & \tau_q^1 \\ \vdots & \vdots & & \vdots \\ \tau_1^{2q-1} & \tau_2^{2q-1} & \cdots & \tau_q^{2q-1} \end{bmatrix}$$

From [5], the required least-squares estimate is

$$\hat{\vec{k}} = -(\mathbf{L}^T \cdot \mathbf{L})^{-1} \cdot \mathbf{L}^T \cdot \vec{m} \quad (15)$$

The error in the model, defined as the amount of mismatch in (8) is now defined as the norm of the vector

$$\vec{e} = \vec{m} - \mathbf{L} \cdot \hat{\vec{k}} \quad (16)$$

This completes the constrained mapping of the moments to a pole-residue representation.

A further enhancement would be to use a weighted least-squares scheme in the second stage of the optimization to ensure that the sum of the residues, m_{-1}, hence the steady-state response, is most accurate. In this case, the estimate is given by

$$\hat{\vec{k}} = -(\mathbf{L}^T \cdot \mathbf{W} \cdot \mathbf{L})^{-1} \cdot \mathbf{L}^T \cdot \mathbf{W} \cdot \vec{m} \qquad (17)$$

where W is a suitable weighting matrix.

Conditioning of matrices is handled by employing frequency scaling [8] and use of the singular value decomposition (SVD) to compute the least-squares fit [7]. In this case,

$$\hat{\vec{x}} = \mathbf{V} \cdot [diag(1/w_j)] \cdot \mathbf{U}^T \cdot \vec{b} \qquad (18)$$

where $\mathbf{L} = \mathbf{U} \cdot [diag(w_j)] \cdot \mathbf{V}^T$ represents the SVD of L.

4 Implementation and Results

As mentioned earlier, calculation of circuit moments for RC interconnects is done very efficiently using a path-tracing technique in [11]. An initial estimation of pole(s) and residue(s) is accomplished using the approach described in [9]. The optimization routines are invoked upon the occurance of a positive pole(s).

For a 4000 node RC circuit, an unconstrained third-order approximation of the unit step response at node 1000 yielded the unstable model (normalized to nanoseconds):

$$\hat{V}(s) = 1 + \frac{7.95}{1+16.2s} + \frac{0.84}{1+1.64s} + \frac{-0.008}{1-1.82s} \qquad (19)$$

Constraining the mapping yielded the stable model:

$$\hat{V}(s) = 1 + \frac{7.93}{1+16.2s} + \frac{0.17}{1+1.51s} + \frac{-0.7}{1+1.78s}. \qquad (20)$$

A graph of the response (20), when plotted versus the output of a circuit simulator (PSPICE) [10], shows highly improved results as compared to the first-order Elmore approximation. (It is noted that the response from PSPICE and (20) are nearly identical.) Further, for a third-order approximation, the algorithm described herein required 1.93 CPU seconds and 708 kilobytes of memory on a Sparcstation 1 while PSPICE required 328.63 CPU seconds and 2368.36 kilobytes on the same machine.

5 Conclusion

A promising approach for overcoming the potential for instability associated with the moment-matching technique in AWE as applied to RC interconnects, has been presented. This approach can further be extended to general linear(ized) RLC models. When used in combination with path tracing, a robust and efficient algorithm is obtained for the problem of accurate RC interconnect delay evaluation.

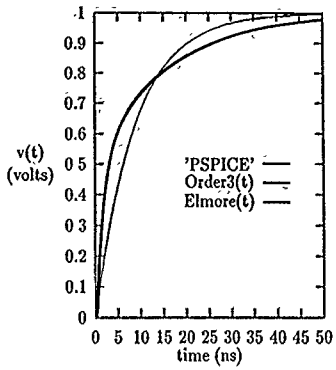


Figure 1. Step response at node 1000 of a 4000 node RC interconnect model

References

[1] M. J. Box. A comparison of several current optimization methods, and the use of transformations in constrained problems. *Comp. J.*, 9, 1966.

[2] R. F. Brown. Model stability in use of moments to estimate pulse transfer functions. *Electron. Lett.*, 7, 1971.

[3] C. Chu and M. Horowitz. Charge sharing models for switch-level simulation. *IEEE Trans. Comp. Aided Design*, (6), 1987.

[4] W. C. Elmore. The transient response of damped linear networks with particular regard to wideband amplifiers. *J. Applied Physics*, 19(1), 1948.

[5] G. H. Hostetter, M. S. Santina, and P. D'Carpio-Montalvo. *Analytical, numerical, and computational methods for science and engineering*. Prentice Hall, Inc., 1991.

[6] X. Huang. *Pade' approximation of linear(ized) circuit responses*. PhD thesis, Carnegie Mellon Univ , Nov 1990.

[7] V. C. Klema and A. J. Laub. The singular value decomposition, its computation and some applications *IEEE Trans Auto. Control*, 25, 1980.

[8] L. T. Pillage. *Asymptotic Waveform Evaluation for Timing Analysis*. PhD thesis, Carnegie Mellon Univ., Apr 1989.

[9] L. T. Pillage and R. A. Rohrer. Asymptotic waveform evaluation for timing analysis. *IEEE Trans. Comp. Aided Design*, (9), 1990.

[10] PSPICE USER'S MANUAL. *Version 4.03*. Microsim Corp., Jan 1990.

[11] C. L. Ratzlaff. A fast algorithm for computing the time moments of RLC circuits. Master's thesis, The Univ of Texas at Austin, May 1991.

[12] J. Rubenstein, P. Penfield, Jr., and M. A Horowitz. Signal delay in RC tree networks. *IEEE Trans. Comp. Aided Design*, (2), 1983.

INTERVAL ARITHMETIC APPLIED TO DIGITAL SIGNAL WAVEFORMS FOR COMPUTATIONALLY EFFICIENT VLSI CIRCUIT VERIFICATION

CHARLES ZUKOWSKI
Member, IEEE
Columbia University
500 W. 120th St. #1311
New York, NY 10027-6699, USA

Abstract - The waveform relaxation algorithm for the electrical simulation of digital circuits is generalized to efficiently handle entire waveform intervals at one time, thus enabling detailed circuit verification that is impractical with non-interval simulators.

I. Introduction

The complexity of digital VLSI circuits makes verification of their detailed operation an extremely difficult problem. In practice, one can simulate only a few subcircuits and situations using current electrical simulators such as SPICE [1]. Even if "exact" simulation could be sped up by orders of magnitude using specialized numerical methods or hardware [2], full verification would still be infeasible due to the large number of cases that must be considered. As a result, for efficient verification we need a simulator that can handle entire sets of waveforms, e.g. waveform intervals, instead of precise waveform estimates. This paper outlines how such intervals can be computed and how they can be used for verification.

Interval arithmetic is used extensively for verifying digital circuits at a high level, e.g. in logic simulation signals are grouped into true and false signal classes. Timing verifiers use estimates of "worst-case" delays to cover a large number of cases with little computation [3]. The use of intervals for verification in switch-level simulation has also recently been introduced [4]. By their very nature, digital circuits have monotonic properties at a high level that make interval arithmetic very practical.

Interval arithmetic has rarely been applied to electrical simulation because it is difficult in general to apply it to the incremental-time solution of ODEs (ordinary differential equations) [5]. Since a small amount of uncertainty can be added at each of the many steps in the computation, acceptable accuracy is often achieved only for small periods of time. In fact, if SPICE were run on an interval computer, it would likely not even be able to distinguish between true and false digital signals by the end of a typical clock period. Fortunately, however, there is an alternate iterative method for solving ODEs called Waveform Relaxation (WR) which performs well on most digital circuits [6] and can be much more efficiently merged with interval arithmetic [7]. WR can lead to efficient interval analysis because it treats signals as entire waveforms, a natural generalization of what is done in logic and timing simulators, and partitions the circuit into small logic blocks. Thus the fundamental high-level monotonic relationships that are built into small digital subcircuits can be exploited.

II. The Waveform Relaxation Algorithm

This section outlines the basic WR method [8]. All WR algorithms start with a guess for the circuit behavior $x^0(\cdot)$, consisting of waveforms for at least a complete set of state variables over some time period $[0,T]$. By simulating subcircuits (e.g., logic blocks) independently over $[0,T]$, this guess is updated and called x^1. The analysis of each subcircuit uses the assumption that connected subcircuits are behaving according to earlier guesses. Continued iteration produces a sequence of behaviors which converges to the actual solution, denoted x^*, for a large and useful class of circuit models [6,7]. A function F can be defined that maps one behavior guess to the next, i.e., $x^{i+1} = F(x^i)$, and WR convergence proofs generally rely on showing that F is contractive in an appropriate norm. The solution x^* can be thought of as a unique "fixed point" of the relaxation mapping function F.

Because signals generally flow primarily in one direction in digital circuits, without any direct feedback other than that delayed by latches, WR can converge very quickly for these circuits when appropriate partitions and schedules are chosen. Since evaluation of F involves incremental-time simulation of each logic block separately, computation time tends to grow only linearly with circuit size. The advantages (efficiency) and disadvantages (complexity and lack of generality) of using WR for "exact" simulation are still the subject of some debate. For verification using waveform intervals, however, it appears to be much easier to build on the WR approach.

III. An Interval Extension of WR

This section illustrates how an interval algorithm for electrical circuit simulation can be derived from a WR algorithm. A set of circuit behaviors, denoted X, is defined by a waveform interval associated with each variable (or union of intervals). Variables in a circuit behavior could consist of node voltages, branch currents, and voltage derivatives. An interval mapping function G, whose domain and range both consist of all possible behavior sets X, is defined in terms of the WR mapping function F as follows:

Def. 1: For all X and any $x \in X$, $F(x) \in G(X)$.

There are many possible G functions, as G is any bound on one of many proposed specific relaxation functions. It is possible to bound F fairly tightly because any F is simply a series of solutions for simple logic blocks. Basic logic blocks have many provably monotonic relationships between inputs and responses, so only endpoints need be

evaluated. In addition, device models can often be rigorously simplified in a manner that drastically reduces computation while only slightly loosening the bounds. G functions have the following properties [7]:

Theorem 1: If $G(X) \subset X$, $x^* \in G(X)$;

Theorem 2: If $x^* \in X$, $x^* \in G(X)$.

These two (simplified) theorems can be used to bound x^* based only on the computation of one or more G functions, possibly repeatedly. The first theorem can be used to verify that a tentative bound on x^* is indeed a bound. If G provides a fairly tight bound on F, it should maintain its contractive property and enable the shrinking of behavior sets. The second theorem can be used to iteratively improve bounds on x^*. Any slack in the function G will limit the tightness of the final bounds on x^*, as will any correlations that are ignored at the block boundaries. The slack arising from such correlations can be managed, however, because the feedback among adjacent blocks is relatively small.

IV. Verification Using Interval Algorithms

Waveform intervals can be used in a number of ways to reduce computation in circuit verification. These uses fall into two main classes, accuracy management and combinatorial compression. These applications are discussed here based on the assumption that a suitable interval waveform simulation algorithm is available.

Accuracy management takes advantage of the fact that uniformly tight bounds for every variable in x^* are not always necessary to verify satisfactory circuit operation [9]. In portions of a combinational logic circuit that surpass their timing specifications, such as logic gates not in critical paths, significant slack can often be tolerated during transitions. Since interval arithmetic allows accuracy to be measured, optimal assignments of accuracy levels can be approached throughout the circuit. Based on monotonic properties of most simple logic blocks, device models can be locally simplified to trade accuracy for computational speed, e.g., using bounding piecewise linear transfer curves.

An "exact" simulator cannot be used for complete verification due to the large number of different cases that must be considered, arising from different combinations of input signals and device parameters. Combinatorial compression involves grouping together classes of inputs into intervals so that computation grows only linearly with circuit size. The use of worst-case delays for input-independent analysis in timing verifiers provides a good example of the basic approach. Waveform interval analysis, though, can achieve similar gains without sacrificing detailed information about waveform shapes. Waveforms can be grouped into classes such as rising, falling, constant-high, and constant-low. Due to the basic monotonicity of logic circuits at the waveform and logic-block level, the extra conservatism that arises from such groupings is generally small.

The potential power of combinatorial compression can be seen by generalizing the switch-level memory verification proposed in [4] to include timing information. For example, to verify the read operation of a single bit in a large memory, all other memory cells can be initialized to the union of two intervals, representing true and false respectively. All the output buses except the one connected to the bit under test will have unknown responses, indicated by very loose bounds that include true and false values. If the bit is still read correctly, however, with a tight interval that guarantees correctness, the circuit has been verified for all possible patterns in the rest of the memory using only a single simulation.

V. Conclusion

Although a full waveform interval simulator has yet to be demonstrated, and many challenges remain to achieve high efficiency, the required basic theory has been developed and experiments have been undertaken that indicate feasibility. Furthermore, such a simulator would provide a powerful new tool for detailed circuit verification, based on its ability to enable accuracy management and combinatorial compression.

Acknowledgements

The author would like to acknowledge support from NSF PYI award MIPS-86-58112.

References

[1] L. Nagel, "SPICE2. A Computer Program to Simulate Semiconductor Circuits," ERL Memo ERL-M520, University of Calif., Berkeley, May 1975.

[2] D. Dumlugol, P. Odent, J. Cocks, and H. DeMan, "Switch-Electrical Segmented Waveform Relaxation for Digital MOS VLSI and its Acceleration on Parallel Computers," IEEE Trans. Comp.-Aided Des., Vol. CAD-6, No. 6, Nov. 1987, pp. 992-1005.

[3] J. Ousterhout, "Crystal. A Timing Analyzer for nMOS VLSI Circuits," *Third CALTECH Conf. on VLSI*, Computer Science Press, March 1983, pp. 57-70.

[4] R. Bryant, "Formal Verification of Memory Circuits by Switch-Level Simulation," IEEE Trans. Comp.-Aided Des., Vol. 10, No. 1, Jan. 1991, pp. 94-102.

[5] R. Moore, *Interval Analysis*, Prentice-Hall, 1966.

[6] E. Lelarasmee, A. Ruehli, A. Sangiovanni-Vincentelli, "The Waveform Relaxation Method for Time-Domain Analysis of Large-Scale Integrated Circuits," IEEE Trans. on Comp.-Aided Des., CAD-1(3), July 1982, pp. 131-145.

[7] C. Zukowski, *The Bounding Approach to VLSI Circuit Simulation*, Kluwer, Boston, 1986.

[8] J. White, A. Sangiovanni-Vincentelli, *Relaxation Techniques for the Simulation of VLSI Circuits*, Kluwer, Boston, 1986.

[9] C. A. Zukowski and G. Dare, "Accuracy Management in VLSI Circuit Simulation," IEEE Symp. on CAS, May 1989, pp. 2027-2031.

Settling-Time Bounds for Switched-Capacitor Networks

Mark N. Seidel
Massachusetts Institute of Technology
Building 36-872
Cambridge, MA 02139 U.S.A.

and

John L. Wyatt, Jr.
Member, IEEE
Massachusetts Institute of Technology
Building 36-872
Cambridge, MA 02139 U.S.A.

Abstract: The Penfield-Rubinstein bounds [1] are extended to cover many switched-capacitor networks. The original formulation in terms of differential equations [2] is modified to include the relevant class of discrete-time difference equations.

1 Introduction

Many algorithms in machine (robot) vision processing can be formulated as minimization problems. Consequently, these algorithms can be implemented using appropriately designed reciprocal resistive networks that relax very quickly to the required solution [3] The appealing features of these networks, namely local computation and communication, make VLSI implementation both feasible and attractive.

Most image processing tasks are accomplished with a large amount of input data (e g , 128×128 intensity values). Power considerations, therefore, require each unit cell to contain relatively large-valued resistors. Since these large values are difficult to achieve consistently and uniformly in standard VLSI, alternatives are sought One such alternative is the use of switched capacitor (SC) resistor equivalents. These equivalents are attractive not only for their low power, but also because of their uniformity, potentially small unit cell size, and dependence on switching frequency.

One drawback of these SC networks is their settling time. Since charge is shared among neighboring capacitors in discrete steps, rather than instantaneously as with continuous-time resistors, the equilibrium charge (and thus voltage) distribution is approached only after some finite number of timesteps. In seeking bounds and estimates for this settling-time waveform, we have extended the Rubinstein-Penfield bounds [1] to the discrete-time case.

2 An Example

There is a large class of SC networks for which bounding expressions can be formed. As an example, consider the SC line shown in Figure 1. There are twenty grounded capacitors, each connected to its two neighbors by either a phase 1 or a phase 2 switch; the two phases alternate in controlling the switches along the line. A single voltage source feeds the line, and, in Figure 1, node k is assumed to be odd. Just for fun, let the capacitors in the line take on the values of the digits of π. That is, let $C_1 = 3$, $C_2 = 1$, $C_3 = 4$, etc.[1] Also, let the voltage on the sixteenth capacitor be the signal of interest.

For the general node, the voltages are for phase 1 (with k odd)

$$v_1\left[n+\frac{1}{2}\right] = v_0 \tag{1}$$

$$v_k\left[n+\frac{1}{2}\right] = v_{k-1}\left[n+\frac{1}{2}\right] = \frac{C_{k-1}v_{k-1}[n]+C_k v_k[n]}{C_{k-1}+C_k} \tag{2}$$

$$v_{20}\left[n+\frac{1}{2}\right] = v_{20}[n] \tag{3}$$

and for phase 2 (with k also odd)

$$v_k[n+1] = v_{k+1}[n+1] = \frac{C_k v_k\left[n+\frac{1}{2}\right]+C_{k+1}v_{k+1}\left[n+\frac{1}{2}\right]}{C_k+C_{k+1}} \tag{4}$$

These equations can be expressed in matrix form, then combined, to yield

$$\mathbf{v}[n+1] = \mathbf{A}\mathbf{v}[n]+\mathbf{b}v_0 \tag{5}$$

[1] To nineteen places, $\pi = 3.14159265358979323\overline{85}$.

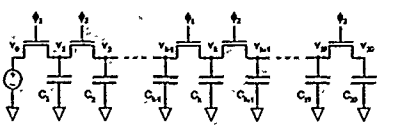


Figure 1. Two-phase SC line of twenty capacitors and switches

Defining $\Delta[n] = \mathbf{v}[n+1]-\mathbf{v}[n]$ and $\mathbf{M} = \mathbf{I}-\mathbf{A}$, the system can be written as

$$\Delta[n] = -\mathbf{M}\mathbf{v}[n]+\mathbf{b}v_0 \tag{6}$$

In this example, M is a nonsingular M-matrix [4] by construction Its inverse, P, exists and has nonnegative entries, and an analysis similar to that found in [2] can be performed. In our example, the three important parameters dependent upon the entries of P are

$$T_{R,16} = 315.7 \tag{7}$$

$$T_{D,16} = 377.3 \tag{8}$$

$$T_P = 408\,3 \tag{9}$$

The bounds, derived below, are shown in Figure 2 The middle signal is the exact response of v_{16} as a function of timesteps, assuming that all capacitor voltages are initially at zero volts. The upper and lower waveforms are the results of the bounding expressions. These bounds are derived from a reduced order model of the system, and represent true bounds. Their advantage is that they are much easier to calculate than the exact expression for large systems.

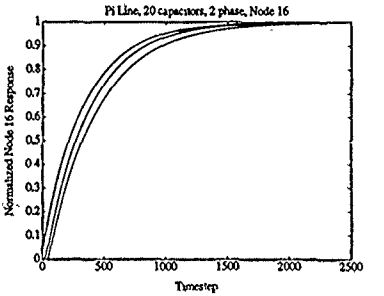


Figure 2. Exact resonse and bounds for node 16 of the SC line.

3 Determining the Bounds

The following derivation is similar to that in [2], altered here to cover the discrete-time case. In general, consider the linear discrete-time system as in equation 5 such that $\mathbf{I}-\mathbf{A}$ is an M-matrix, where $\mathbf{b} > 0$ (that is, every element is nonnegative and $\mathbf{b} \neq 0$) and v_0 denotes the distinguished component of v that serves as the system output. Let v_0 be a positive step input. Define $\Delta[n]$, M, and P as before, along with the equilibrium (final) state value

$$\mathbf{v}_{eq} \triangleq \mathbf{P}\mathbf{b}v_0 \tag{10}$$

The system can now be rewritten as

$$\mathbf{P}\Delta[n] = \mathbf{v}_{eq} - \mathbf{v}[n] \tag{11}$$

It can be shown that the zero-state step response of the distinguished component v_i of the system state is a monotonic nondecreasing function of timesteps with final value v_{ieq}. The function $n_{min}(v_i^*)$ is defined as the minimum possible number of steps necessary to achieve $v_i[n] \geq v_i^*$, while $n_{max}(v_i^*)$ is defined as the maximum possible number of steps before the system output $v_i[n]$ is forced past the value v_i^*. An output level v_i^* will be reached in no less than n_{min} steps and crossed in no more than n_{max} steps.

We construct a reduced order model with two states defined by

$$g_i[n] \triangleq \frac{v_{ieq} - v_i[n]}{v_{ieq}} \tag{12}$$

$$f_i[n] \triangleq \sum_{k=n}^{\infty} g_i[k] \tag{13}$$

that obeys

$$(f_i[0], g_i[0]) = (T_{Di}, 1) \tag{14}$$

$$f_i[n+1] = f_i[n] - g_i[n] \tag{15}$$

$$g_i[n+1] \leq g_i[n] \tag{16}$$

$$g_i[n]T_{Ri} \leq f_i[n] \leq g_i[n]T_P \tag{17}$$

$$v_i[n] = v_{ieq}(1 - g_i[n]) \tag{18}$$

where the three system parameters, T_{Ri}, T_{Di}, and T_P, are defined in [2]. The equalities in equations 14-18 follow from the definitions, and the inequalities follow from the fact that I − A is an M-matrix.

The reduced order model's equations can be graphically interpreted as shown in Figure 3.

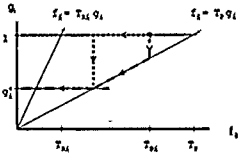


Figure 3. Reduced order model statespace geometry. (Dotted line. n_{min} path. Dashed line. n_{max} path.)

At every timestep, the state jumps to a new point as dictated by the reduced order model (the next value of f_i is specified, and the next value of g_i is bounded); the starting point is at $(T_{Di}, 1)$. The goal is either to reach the target line ($g_i = g_i^*$) as quickly as possible (n_{min}), or to delay going past it as long as possible (n_{max}). The lines $f_i = T_{Ri}g_i$ and $f_i = T_P g_i$ are the upper and lower boundary lines, respectively.

The procedure for obtaining the minimum number of timesteps from the initial point to a point on or below the target line is to follow the "highest" path through (f_i, g_i)-space. This path consists of traversing the line $g_i = g_i[0]$, then traversing the upper boundary, always in the direction of lower f_i values. This path is the fastest way to move towards lower f_i values, since each f_i step is as large as possible. Of course, the final jump to or below the target line is taken as early as possible. By following this methodology, the n_{min} value arrived at is the minimum possible number of timesteps necessary to move on or below the target line.

The procedure for obtaining the maximum number of timesteps from the initial point to a point below the target line is to follow the "lowest" path through (f_i, g_i)-space. This path consists of traversing the line $f_i = f_i[0]$, then traversing the lower boundary line, and finally traversing the target line, always in the direction of lower f_i values. This path is the slowest way to move towards lower f_i values, since each f_i step is as small as possible. Of course, the final jump below the target line is taken as late as possible. By following this methodology, the n_{max} value arrived at is the maximum possible number of timesteps such that its state and every subsequent state is forced below the target line.

The discrete-time bounds have the following form:

$n_{min}(v_i^*)$	v_i^* range
0	$v_i^* = 0$
$\lceil \max(1, T_{Di} - T_P g_i^*) \rceil$	$0 < \frac{v_i^*}{v_{ieq}} \leq \frac{T_P - T_{Di} + n_a}{T_P}$
$n_a + \left\lceil \frac{\log\left(\frac{T_P g_i^*}{T_{Di} - n_a}\right)}{\log\left(1 - \frac{1}{T_P}\right)} \right\rceil$	$\frac{T_P - T_{Di} + n_a}{T_P} < \frac{v_i^*}{v_{ieq}} < 1$

$n_{max}(v_i^*)$	v_i^* range
$2 + \left\lfloor \frac{T_{Di} - 1}{g_i^*} - T_{Ri} \right\rfloor$	$0 \leq \frac{v_i^*}{v_{ieq}} < \frac{T_P - T_{Di} + 1}{T_P}$
$2 + n_b + \left\lfloor \left(\frac{T_{Di} - 1}{g_i^*}\right)\left(1 - \frac{1}{T_P}\right)^{n_b} - T_{Ri} \right\rfloor$	$\frac{T_P - T_{Di} + 1}{T_P} \leq \frac{v_i^*}{v_{ieq}} < 1$

where

$$g_i^* = \frac{v_{ieq} - v_i^*}{v_{ieq}} \tag{19}$$

$$n_a = 1 + \lfloor T_{Di} - T_{Ri} \rfloor \tag{20}$$

$$n_b = 1 + \left\lfloor \frac{\log\left(\frac{T_P g_i^*}{T_{Di} - 1}\right)}{\log\left(1 - \frac{1}{T_P}\right)} \right\rfloor \tag{21}$$

The symbols $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor and ceiling functions.

4 Conclusions

The bounds are known to apply to a large class of SC networks, namely those containing an arbitrary number of grounded capacitors and grounded independent voltage sources, along with an arbitrary number of switches connecting any of the elements. The class of networks is sure to be more extensive, but the enlarged boundaries have not been fully investigated. Tightening the bounds is also an important topic, as is studying under what conditions the P matrix may be stuffed by inspection (to avoid calculating a matrix inverse). It is known that the P matrix can be stuffed by inspection for a certain class of SC trees.

References

[1] J. Rubinstein, P. Penfield, and M. A. Horowitz, "Signal Delay in RC Tree Networks", *IEEE Trans. Comp.-Aided Des.*, vol. CAD-2, pp. 202-211, July 1983.

[2] J. L. Wyatt, Jr., C. A. Zukowski, and P. Penfield, Jr., "Step Response Bounds for Systems Described by M matrices, with Application to Timing Analysis of Digital MOS Circuits", in *Proc 24th IEEE Conf. Dec. and Cont.*, (Fort Lauderdale, FL), pp 1552-1557, Dec. 11-15, 1985.

[3] A. Lumsdaine, J Wyatt, and I. Elfadel, "Parallel Distributed Networks for Image Smoothing and Segmentation in Analog VLSI", in *Proc. 28th IEEE Conf. Dec. and Cont*, (Tampa, FL), Dec. 1989.

[4] A. Berman, Nonnegative Matrices in the Mathematical Sciences, Academic Press, New York, 1979.

Overlapping Domain Decomposition Method for a Unilateral Boundary Value Problem

Yu. A. Kuznetsov
Department of Numerical Mathematics
Academy of Sciences of the USSR
Moscow, USSR

P. Neittaanmäki
Department of Mathematics
University of Jyväskylä
PL 35, 40351 Jyväskylä, Finland

Abstract. In this paper overlapping domain decomposition methods are applied to the numerical solution of Poisson equation with Dirichlet-Signorini boundary condition.

1. Introduction

In this paper we describe the numerical solution by overlapping domain decomposition methods of grid variational problems arising from finite element approximations of the simplified Dirichlet-Signorini problem for Poisson equation. Probably, the paper [1] was the first, where overlapping domain decomposition methods were applied to the solving of variational inequalities. The convergence proof of our methods is based on the finite dimensional approach proposed and developed in [3] for linear variational problems.

We consider the problem

$$\begin{aligned} &-\Delta u = f \text{ in } R^2, \\ &u = 0 \text{ on } \Gamma_0 \ (\neq 0), \\ &u \geq 0,\ \frac{\partial}{\partial n} u \geq 0,\ u \cdot \frac{\partial}{\partial n} u = 0 \\ &\qquad \text{a.e. on } \Gamma_1 = \partial\Omega \setminus \overline{\Gamma}_0, \end{aligned} \tag{1.1}$$

where $\frac{\partial}{\partial n}$ denotes the outer normal derivation to $\partial\Omega$. It is well known that (1.1) can be formulated as a minimization problem:

$$\text{Find } u \in K : J(u) = \min_{v \in K} J(v) =$$

$$\frac{1}{2}\int_\Omega \nabla v \nabla v \, dx - \int_\Omega f v \, dx \equiv \frac{1}{2} a(v,v) - (f,v), \tag{1.2}$$

where $K = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_0 \text{ and } v \geq 0 \text{ on} \Gamma_1\}$. The equivalent formulation of the problem (1.1), (1.2) in terms of variational inequality reads

$$\text{Find } u \in K : a(u, v-u) \geq (f, v-u) \ \forall v \in K. \tag{1.3}$$

2. Finite element approximation

Let $\mathcal{T}_h = \{e_i^h\}$ be a triangular covering of Ω under the typical assumptions like that the set $\Gamma_0 \cap \overline{\Gamma}_1$ belongs to the set of vertices of triangle elements e_i^h. We denote a standart piecewise linear finite element subspace of V related with $\mathcal{T}_h$ by V_h and consider the finite element approximation of the Dirichlet-Signorini problem (1.1) in the form of minimization task:

$$\text{Find } u_h \in K_h : J(u_h) = \min_{v \in K_h} J(v) \tag{2.1}$$

where $K_h = V_h \cap K$.

The equivalent formulation of (2.1) in the terms of variational inequalities is defined by:

$$\text{Find } u_h \in K_h : a(u_h, v-u_h) \geq (f, v-u_h) \ \forall v \in K_h. \tag{2.2}$$

3. Overlapping domain decomposition method

G_h is said to be a grid subdomain of $\mathcal{T}_h$ (of $\overline{\Omega}$), if G_h is a union of triangle elements e_i^h. The interior of G_h is denoted by $\overset{0}{G}_h$ and its boundary is denoted by ∂G_h. The boundary of G_h consists of two subsets: $\Gamma_0^{(G)}$ denotes the closure of the intersection of ∂G_h with $\Omega \cup \Gamma_0$, i.e. $\Gamma_0^{(G)} = \overline{\partial G_h \cap (\Omega \cup \Gamma_0)}$, and $\Gamma_1^{(G)}$ denotes the remaing part of ∂G_h, i.e. $\Gamma_1^{(G)}$ is the interior of $\partial G_h \cap \Gamma_1$.

Let a grid subdomain $G_h \subset \mathcal{T}_h$ and a finite element function $v_h \in K_h$ be given. If $\partial G_h = \Gamma_0^{(G)}$ we shall associate with the pair $(G_h; v_h)$ the only subset $\overset{0}{K}_h(G; v_h) \subset K_h$, which is defined by:

$$\overset{0}{K}_h(G; v_h) = \{w_h \in K_h \mid w_h\big|_{\overline{\Omega} \setminus \overset{0}{G}_h} = v_h\}. \tag{3.1}$$

If $\Gamma_1^{(G)} \neq \emptyset$, we shall associate with the pair $(G_h; v_h)$ either the subset $\overset{0}{K}_h(G; v_h)$ of K_h or the subset $\overset{1}{K}_h(G; v_h)$

of K_h which is defined by:

$$\overset{1}{K}_h(G; v_h) = \{w_h \in K_h \mid w_h\big|_{\overline{\Omega}\setminus(G_h \cup \overset{0}{\Gamma}_1^{(G)})} = v_h\}. \tag{3.2}$$

Note that both subsets $\overset{0}{K}_h(G; v_h)$ and $\overset{1}{K}_h(G; v_h)$ are closed and convex. Thus, we associate with the pair $(G_h; v_h)$ the subset $K_h(G; v_h) \subset K_h$, which is equal to either $\overset{0}{K}_h(G; v_h)$ or $\overset{1}{K}_h(G; v_h)$.

Let us consider the following variational subdomain problem: Find $w_h \in K_h(G; v_h)$ such that

$$J(w_h) = \min_{\varphi_h \in K_h(G; v_h)} J(\varphi_h), \tag{3.3}$$

which is equivalent to the variational inequality: Find $w_h \in K_h(G; v_h)$ such that

$$a(w_h, \varphi_h - w_h) \geq (f; \varphi_h - w_h) \forall \varphi_h \in K_h(G; v_h). \tag{3.4}$$

It is obviously, that the latter problem has a unique solution.

The problems (3.3), (3.4) result

$$\begin{aligned} J(v_h) &= J(w_h + (v_h - w_h)) \\ &= J(w_h) + a(v_h - w_h, v_h - w_h) + \\ &\quad 2\{a(w_h, v_h - w_h) - (f, v_h - w_h)\} \\ &\geq J(w_h) + a(v_h - w_h, v_h - w_h), \end{aligned}$$

which leads to the following conclusion.

LEMMA 3.1.

$$J(w_h) \leq J(v_h)$$

for any $v_h \in K_h$ *and* $J(w_h) = J(v_h)$ *iff* $w_h = v_h$.

Let us assume that a grid subdomain G_h and a function f are fixed as well as a triangular covering $\mathcal{T}_h$. We define the operator $T_G : K_h \mapsto K_h$ such that for any $v_h \in K_h$ the solution function w_h of (3.3), (3.4) is given by formula

$$w^h = T_G(v_h). \tag{3.5}$$

LEMMA 3.2. *The operator* $T_G : K_h \mapsto K_h$ *is continuous.*

Let m be a positive integer and let $\mathcal{T}_h$ be decomposed into m grid subdomains $G_j = G_h^{(j)}$, i.e. $\mathcal{T}_h = \bigcup_{j=1}^m G_j$. We associate with every subdomain G_j a subset $K_h^{(j)}$ of K_h, which is equal to either $\overset{0}{K}_h(G)$ or $\overset{1}{K}_h(G)$ where $\overset{0}{K}_h(G)$ and $\overset{1}{K}_h(G)$ are defined as in (3.1) and (3.2). The latter case means that with every index j we associate the pair $(G_j; K_h^{(j)})$, $j = \overline{1, m}$.

We shall require that the following two conditions are satisfied:

(A1) every grid node $x_i \in \Omega_h$ belongs to at least one $\overset{0}{G}_j$;

(A2) every grid node $x_i \in \Gamma_{1,h}$ belongs to at leas one $\Gamma_1^{(j)} \equiv \Gamma_1^{(G_j)}$.

The assumption (A1) means that every grid node from Ω_h belongs to the interior of at least one subdomain G_j. In assumption (A2) we require that every grid node from $\Gamma_{1,h}$ belongs to the interior of at least one $\Gamma_1^{(j)}$, i.e. in particular for this index j we have $K_h^{(j)} = \overset{1}{K}_h(G_j)$.

According to above definitions and assumptions we can formulate the following iterative domain decomposition procedure.

Algorithm. Let $u^0 \in K_h$ be given. Solve successively for $n = 0, 1, \ldots$ and for $j = 1, \ldots, m$ the following subproblems: Find $u_h^{n+\frac{j}{m}} \in K_h(G_j; u_h^{n+\frac{j-1}{m}})$ such that

$$a\left(u_h^{n+\frac{j}{m}}, \varphi_h - u_h^{n+\frac{j}{m}}\right) \geq \left(f, \varphi_h - u_h^{n+\frac{j}{m}}\right) \quad \forall \varphi_n \in K_h(G_j; u_h^{n+\frac{j-1}{m}}) \tag{3.5}$$

THEOREM 3.1. *Under the assumptions (A1) and (A2) the sequence* $\{u_h^{n+\frac{j}{m}}\}_{n=0}^{\infty}$ *convergence for every* j $(1 \leq j \leq m)$ *to the solution function* u_h *of the problems (2.1), (2.2) for any* $u_h^0 \in K_h$.

PROOF: See [2].

4. EXAMPLES

EXAMPLE 4.1: Let Ω be a rectangle and r be a positive integer. To this end we set $G_1 = \Omega_1$ and associate $\overset{0}{K}(\Omega)$ with G_1 and then choose $G_{2i} = \tilde{G}_1$ and

$G_{2i+1} = \bar{G}_2$, $i = 1, 2, \ldots, r$ as shown in Fig. 1. We associate $K(G_j)$ with G_j for $j = 2, \ldots, 2r+1$. Because under the assumptions made the conditions of Theorem 3.1 are satisfied the corresponding iterative method (3.5) is convergent.

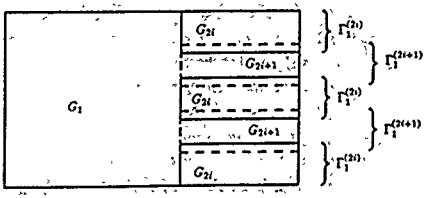


Fig. 1. Decomposition of a rectangle Ω into subdomains $G_1 = \Omega$ and G_j, $j = \overline{2, 2r+1}$

Numerical experiments. Let $\Omega = (0,1) \times (0,1)$ be the unit square, $\Gamma_1 = \{(x_1, x_2) \mid x_1 = 1, \quad x_2 \in (0,1)\}$ be a side of this square and let $f = f(x_1, x_2)$ be a given piecewise constant function such that

$$f = \begin{cases} f_1 = \text{const} < 0, & x_2 \in (0, \frac{1}{2}), \\ f_2 = \text{const} > 0, & x_2 \in (\frac{1}{2}, 1). \end{cases}$$

Partition Ω into subdomains G_j, $j = \overline{1, 2r+1}$ which are constructed as shown in Fig. 1, where r is a positive integer. It follows, that we set $G_1 = \Omega$ and G_j are multiconnected domains consisting of $t+1$ equal rectangles for even values of j and t equal rectangles for odd values of j, where t is a positive integer, $j = \overline{2, 2r+1}$. For numerical experiment the separate rectangles of subdomains G_j were chosen with sides equal to $(s+1)h$ and $2h$ for x_1 and x_2 axes respectively, where s is a positive integer.

Let k_ε be a number of steps of the iterative method (3.5) under the assumptions made, which are required to get the finite element solution u_h on Γ_1 with a prescribed accuracy ε. It is obvious, that k_ε depends on values of s and m. In Tables 1 and 2 we present the dependence of k_ε of s and m, which was established by numerical experiments for $\varepsilon = 10^{-5}$ and suitable criteria for the stopping on the iterative method. We set $h = \frac{1}{32}$ in Table 1 and $h = \frac{1}{64}$ in Table 2.

s \ m	1	2	3	4
3	64	42	29	21
7	64	42	29	20
15	64	42	29	20

Table 1.

s \ m	1	2	3	4
3	117	78	54	40
7	117	78	54	39
15	117	78	54	39

Table 2.

The results of Table 1 and 2 show that the iterative method convergences quite rapidly and the rate of convergence becomes better for bigger values of r. At the same time the rate of convergence is practically independent of s. Probably, this is a partial property of the concret problem under consideration. For further numerical tests see [2].

References

[1] P.L. Lions, *On the Schwarz Alternating Method I*, In "Domain Decomposition Methods for PDE's", R. Glowinski, G.H. Golub, G.A. Meurant and J. Periaux (eds.), SIAM, Philadelphia (1988), 2–42.

[2] Yu.A. Kuznetsov and P. Neittaanmäki, *Overlapping Domain Decomposition Method for Simplified Dirichlet-Signorini Problem*, Preprint 126, University of Jyväskylä, Dept. Math. (1991).

[3] G.I. Marchouk and Yu.A. Kuznetsov, *Methodes iteratives et fonctionelles quadratiques*, Methods Mathematiques de L'Informatique 4; Sur les methodes numeriques en sciences, physiques et economiques, J.L. Lions and G.I. Marchouk (eds.), Dunod, Paris (1987), 3–132.

MODELLING OF InGaAs HEMT

H. HAPPY, O. PRIBETICH, G. DAMBRINE
J. ALAMKAN, Y. CORDIER, A. CAPPY

Centre Hyperfréquences et Semiconducteurs U.A. 287 CNRS-Bât P4
Université des Sciences et Techniques de Lille Flandres Artois
59655 Villeneuve d'Ascq Cedex - France

Abstract: *HELENA, a new software for the modelling of pseudomorphic HEMT and more generally for any kind of HEMTs is presented. This software provides the DC, AC and noise properties in the cm and mm wave range. It is very fast, easy to use and needs only personal computer.*

I-INTRODUCTION.

For the design of MMICs, and more generally for the optimization of circuits using HEMTs , the knowledge of the high frequency parameters including the noise parameters constitutes a key point. Since measurements on test device are difficult (especially for the noise parameters), a theoretical modelling is very well suited if it is in good agreement with experiments. For this purpose a new HEMT software called HELENA for Hemt ELEctrical properties and Noise Analysis has been developped.

The flow chart of HELENA is given in figure 1

It is divided in three blocks :

- The layer analysis
- The device performance modeling
- The results display

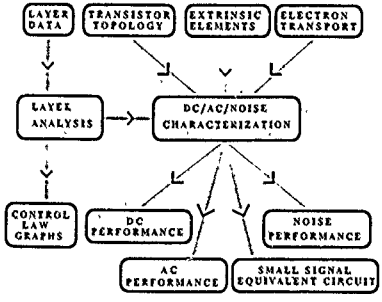


Fig.1:HELENA flow shart

II-THE LAYER ANALYSIS

The purpose of this routine is to obtain the charge control law by a schottky barrier of any HEMT active layer . The input parameters are the thickness, doping (bulk or planar) and composition (%In, %Al) of each layer constituting the active layer.

This modeling is based on a simplified selfconsistent resolution of Schrodinger and Poisson's equations taking the strain effects into account . As the matter of fact, J. ALAMKAN (1) has shown that the two first energy subbands are related to the sheet carrier density in 2DEG by the relation

$$E_i = A_i + B_i . N_S \quad [1]$$

where A_i and B_i are two constants and N_S is the sheet carrier density. This result allows us to write a simplified charge control law for electrons in the 2DEG . The electron supplying layer for which quantum effects are negligible is classically described using Fermi statistics. Layer analysis provides for each layer the sheet carrier density (ionized donnor and free carrier density), and associated capacitances as a function of the gate voltage. Figure 2 shows the charge control law in terms of capacitances in the 2DEG and in the electron supplying layer for the layer shown fig. 2.

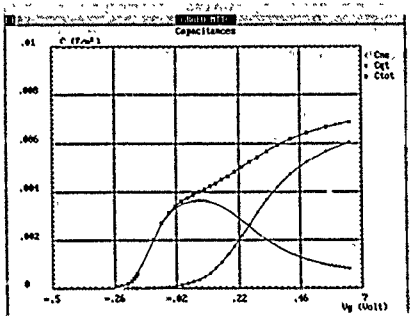


fig 2: Charge control law

III-THE DEVICE MODELING

The device modeling is performed using a quasi 2D approach (2),(3). In this approach the charge control law (in terms of sheet carrier density) is associated with an electron transport law to give the DC, AC and Noise properties. The input parameters of this routine are the transistor topology , the charge control law given by the layer analysis previously described, the electron transport law and the extrinsic parameters of the small signal equivalent circuit. For the electron transport law, both simple v(E) relationship (suitable for fitting experimental data) or hydrodynamic equation (suitable for device performances predictions) can be used. A

realistic structure including gate recess as well as surface potential and carrier injection into the buffer is taken into account.

III-a : *DC AND AC CHARACTERISTICS*

For given V_{gs} and I_{ds}, electron transport and Poisson's equation are combined to give a quadratic equation (4). Solving this equation at each step from source to drain gives the quantities of physical interest in the channel: Electric field, electron velocity, electron energy... Drain to source voltage is obtained by integrating the electric field from source to drain. The DC characteristics $I_{ds}(V_{gs}, V_{ds})$ are then computed (see Fig. 3).

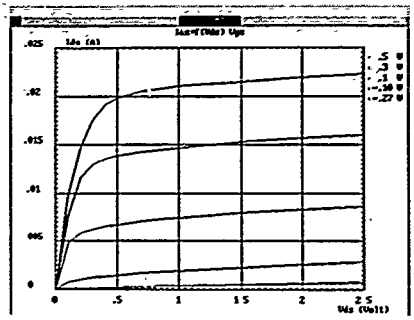


fig 3: DC characteristics

At each DC point, the AC performance (S-parameters or small signal equivalent circuit) is computed using the active line approach (5).. A feature of this approach is to provide not only the main small signal parameters (G_m, G_d, C_{gs}, C_{gd}) but also R_i, Rgd, τ, C_{ds} which are very important for the determination of high frequency device behaviour (see Fig.3).

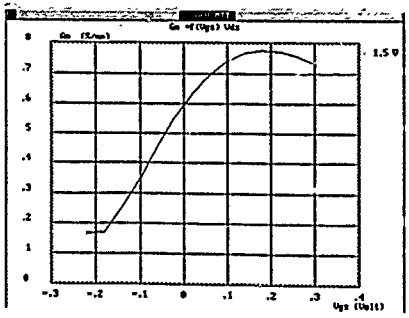


fig4 Transconductance versus gate-to-source voltage

III-b: *NOISE MODELING*

For each DC point, the four noise parameters F_{min}, R_n, G_n, Γ_{opt} as well as associated gain (G_{ass}) and maximum available gain (MAG) are computed using the impedance field method. The impedance field is deduced from the active line following the method described in (5). The intrinsic device is first considered and after that the extrinsic elements (inductance, capacitance, resistance) are introduced

Figure 5 shows the frequency variation of Fmin for a 0.3 micron gate length pseudomorphic HEMT.

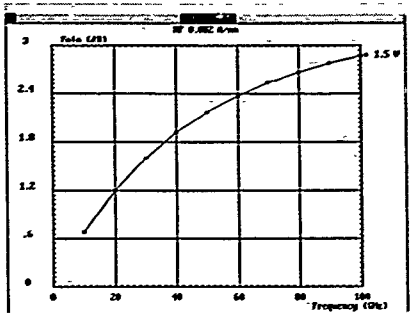


fig 5 Minimum noise figure versus frequency

REFERENCES

(1) J. Alamkan, H. Happy, Y. Cordier, A. Cappy "Modeling of pseudomorphic AlGaAs/GaInAs/GaAs layers using selfconsistent approach", European transactions on telecommunications and related technologies, Vol. I No 4July-August 1990.

(2) A. Cappy, A. Vanoverschelde, M Schortgen, C. Versnaeyen, G. Salmer "Noise Modeling in submicrometer Gate Two dimensional Electron-Gas Field Effect Transistor", IEEE Trans. on Electron Devices, Vol. 32, No. 12. February 1985.

(3) Snowden, C.M. and Pantoja R.R., "GaAs-two-dimensional MESFET Simulation for CAD", IEEE Trans. on Electron Devices, February 1989, pp 1564-1574.

(4) B. Carnez, A. Cappy,, R. Fauquembergue, E. Constant, G. Salmer, "Noise modeling in Submicrometer-Gate FET's", IEEE Trans. on electron devices, Vol. ED-28, No 7, July 1981

(5) A. Cappy, W. Heinrich, "High-Frequency FET Noise Performance: A New Approach", IEEE Trans. on Electron Devices, Vol. 36, No. 1. February 1990, pp 403-409.

SIMULATION OF InGaAs FETs

J.Mehegan, K.McCarthy, N.Cordero, W.Kelly†, C.Lyden†

National Microelectronics Research Centre,
University College, Cork, Ireland.

†Farran Technology Ltd.,
Cork, Ireland.

Abstract – We describe a two dimensional drift-diffusion model used to simulate InGaAs JFETs and Barrier Enhanced MESFETs. We also describe a one dimensional model used to investigate the multi-layer structure of BE-MESFETs.

I. INTRODUCTION

The electron transport properties place InGaAs-based field effect transistors (FETs) amongst the most promising devices in the design of opto-electronic integrated circuits for high speed optical fibre communications systems [1]. Standard MESFET technology cannot be used in the fabrication of these FETs as the Schottky barrier height on n-doped InGaAs is too low ($\sim 0.2eV$) [2]. Hence some other method of achieving transistor action is necessary. For a simple n-channel device two different structures are:

(1) Junction FET (JFET) - a p-layer placed under an ohmic contact gate, achieving modulation through gate voltage control of the depletion width of the p-n junction formed under the gate.

(2) Barrier Enhanced MESFET (BE-MESFET) - a Schottky contact to an intermediate layer of higher barrier height and bandgap. This higher Schottky barrier and conduction band discontinuity between the gate layer and the active layer gives rise to an effective barrier on the channel. Modulation is then achieved in a similar manner to the "classical" MESFET.

Numerical simulation of each of these options is necessary in the effort to develop and optimise this technology. Simulation over the entire device is performed using a 2D drift diffusion model described below. The details of the layer structure in BE-MESFETs are more accurately investigated using a 1D model. This involves solving the Poisson equation over a cross section from the gate contact into the bulk, yielding the band edge profiles and electron distributions across the layers.

II. THE MODELS

Drift-Diffusion Model

The two dimensional model is based on a previously reported MESFET simulator [3]. It uses the semi-classical steady state description of current flow, solving for the Poisson and current continuity equations

$$\nabla(\epsilon\nabla\phi) = q(n - p + N_A - N_D) \tag{1}$$

$$\nabla[n\mu_n\nabla\phi - \nabla(D_n n)] = G \tag{2}$$

$$\nabla[p\mu_p\nabla\phi + \nabla(D_p p)] = -G \tag{3}$$

where G is the generation-recombination (G-R) rate. G-R is modelled as Schockley-Read-Hall type

$$G = \frac{n_i^2 - pn}{\tau_n(p + n_i) + \tau_p(n + n_i)} \tag{4}$$

though other generation models can easily be included.

The model uses the finite element method allowing recessed gate structures and many other non rectangular geometries to be simulated.

The dependence of electron mobility in InGaAs with electric field is modelled by

$$\mu_n E = \left(\frac{\mu_0^4 E^4}{1 + (\mu_0 E / v_{sat})^4}\right)^{1/4} \tag{5}$$

with $\mu_0 = 5000cm^2/Vs$ and $v_{sat} = 1 \times 10^7 cm/s$. For holes a constant mobility $\mu_p = 200cm^2/Vs$ is used. This ignores overshoot effects to bring the simulations into line with measured results on real devices [4]. The Einstein relation between diffusivity and mobility is assumed for both carrier species

$$D_{n,p} = \frac{kT}{q}\mu_{n,p} \tag{6}$$

In simulating JFETs the full two carrier model is needed. It is possible to consider only electron conduction with the inclusion of stationary hole charge to define the p-n junction. However this is inadequate to investigate gate leakage current, an important phenomenon in these devices.

For BE-MESFETs we need consider electron conduction only. The effect of the conduction band offsets between the different layers in these devices is included by solving for an effective potential ϕ^* in the current continuity equation

$$\phi^* = \phi + \chi + \frac{kT}{q}ln(N_c) \tag{7}$$

where χ is the electron affinity of the different layers and N_c the conduction band density of states. This approach ignores tunnelling, thermionic and quantum effects, but is valid for BE-MESFET simulation where such effects are negligible. A more appropriate model is needed where transport between layers may be a predominant effect (e.g. HEMTs).

One Dimensional Model

Solution of the Poisson equation over the 1D cross section requires an accurate description of the dependence of electron concentration on potential across the layers. In many multi-layer FET structures the electron "gas" is degenerate and its energies often quantised. We ignore quantisation for BE-MESFETs but do account for degeneracy through the use of Fermi-Dirac statistics. This gives

$$n(z) = \frac{\sqrt{\pi}}{2} N_c(z) F_{1/2}[(E_F - E_c(z))/kT] \tag{8}$$

as the relation between conduction band edge $E_c(z)$ and electron concentration. $F_{1/2}$ is the Fermi-Dirac integral of order $\frac{1}{2}$ for which Rational Chebyshev approximations are known [5]. The fermi level E_F is the same across the different layers, assuming equilibrium between the layers, thus acting as a natural zero level for energy. $E_c(z)$ is related to the potential ϕ by

$$E_c(z) = -\phi(z) + \Delta E_c(z) \tag{9}$$

where $\Delta E_c(z)$ describes the band offset between different layers in the program with reference to the gate layer.

Discretization and solution of the resulting non-linear Poisson equation are easily performed over the 1D mesh. Simulation yields the electron concentration and band edge profile.

III. NUMERICAL RESULTS

Simulation of the InGaAs JFET shown in fig.1 was performed using the full two carrier solver. This was to investigate reverse bias gate leakage current Igs-Vgs for different values of Vds (fig.2). The recombination times $\tau_n = \tau_p = 5 \times 10^{-4}s$ were used in (4).

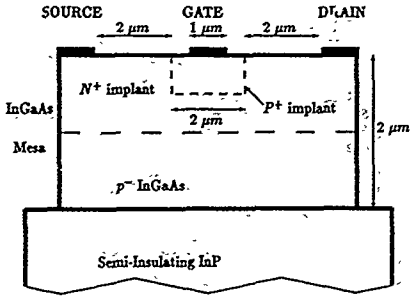


Figure 1: InGaAs JFET cross section.

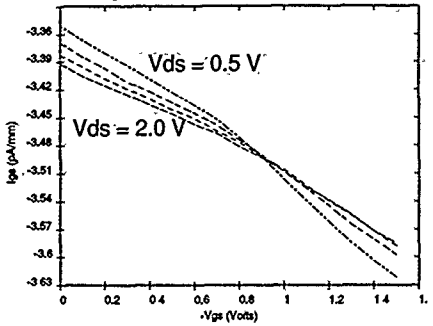


Figure 2: Igs/Vgs for Vds = 0 5, 1.0, 1.5 and 2.0 V

A BE-MESFET with InAlAs gate and buffer layers was also studied using both 1D and 2D simulators. The modulation of the device is clearly seen from the band edge profiles in fig 3. These also show the extent of the barrier to reverse bias gate leakage current, justifying the use of (7). The *Ids-Vds* DC characteristics are shown in fig.4. The gate bias dependence of both the sheet charge in the channel, obtained using the 1D solver, and the current *Ids* are shown in fig.5. It can be clearly seen that gate voltage control over the sheet charge in the channel is the dominant effect in determining the transconductance of the device.

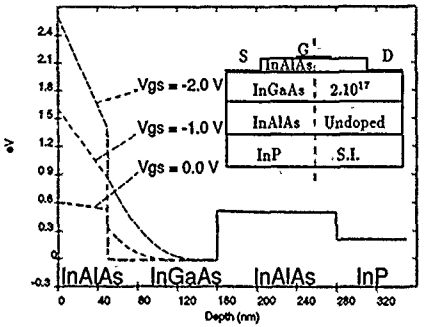


Figure 3: Band Edge profile

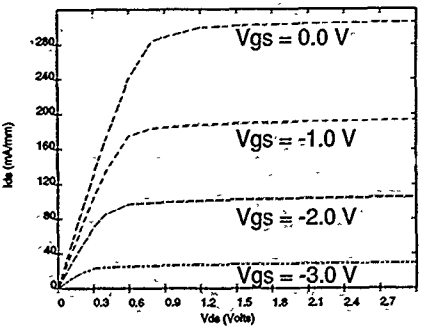


Figure 4: Ids/Vds curves for different Vgs

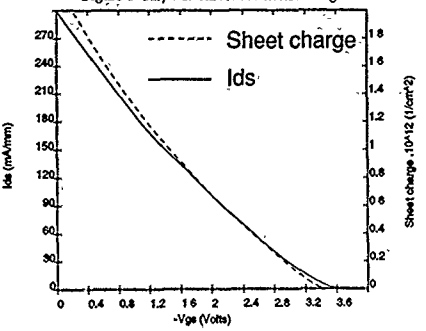


Figure 5: Ids and Sheet charge dependence with Vgs

IV. CONCLUSIONS

We have described a full 2D drift-diffusion model and demonstrated how it is implemented to simulate InGaAs based JFETs and BE-MESFETs. We have also shown how a more accurate 1D model is used to describe the band edge structure and the charge distribution in BE-MESFETs.

ACKNOWLEDGEMENTS

This work was partly supported by the E.C. through the ESPRIT project no. 2035. The authors would also like to thank S Rosser and M.Agnew of STL (Harlow) for supplying material parameters and other useful communications.

REFERENCES

[1] K. Heime, *InGaAs Field-Effect Transistors*, Research Studies Press Ltd., 1989.
[2] K. Kajiyama, Y. Mizushima, S.Sahata, *Appl Phys. Lett.*, vol. 23, p. 458, 1973.
[3] C. Lyden, W. Kelly, J. Campbell, S. Eivers, *Second International Conference Simulation of Semiconductor Devices and Processes*, vol. 2, p. 508, University College of Swansea, 1986.
[4] N. Cordero, K. McCarthy, C. Lyden, W. Kelly, *to be published.*
[5] W.J. Cody, H.C. Thacher, Jr., *Math. Comput.* 21 30, 1967.

TRANSIENT THERMAL MODELLING OF GALLIUM ARSENIDE BIPOLAR HETEROJUNCTION POWER TRANSISTORS

P.W.WEBB
School of Electronic and Electrical Engineering
University of Birmingham UK.

and

I.A.D.RUSSELL
School of Electronic and Electrical Engineering
University of Birmingham

Abstract The optimal thermal design of microwave power devices is complicated because the electrical design dictates that the layout of the structure should be as small as possible and the thermal considerations require that the device should be large for minimum thermal resistance. If such a device is to be used under transient conditions with some known duty cycle the requirement for the device's physical size will be to some extent be reduced but the optimal design problem becomes more complicated. This presentation outlines some of the numerical simulation problems which exist for the transient thermal design of heterojunction bipolar power transistors (HBJT's) and a method of simplifying the design process is suggested.

1. INTRODUCTION.

The question of physical size in the design of microwave devices is always extremely important because of the minimisation of parasitic components in the electronic equivalent circuit and matching its terminal impedance. Generally the thermal design dictates that the device should be as large as possible in area and as thin as possible to reduce its thermal resistance, the benefits being greater reliability and more predictable electrical characteristics. The electrical characteristic will be a function of temperature whether the device is being used under CW steady state conditions or pulsed according to some duty cycle. One problem that exists with pulsed operation is the change in the electrical characteristics during the time that the devices is operating. The thermal design of a microwave device to be used under transient or pulsed conditions is an interesting one as it should be possible to make the device smaller then a device designed for CW operation giving improved bandwidth, efficiency and power output. Current designs using HBJT's consist of a variety of materials with temperature dependant properties requiring 3D simulation. The computer time required to perform these transient simulations can become embarrassingly long, and this paper seeks to outline the problem and suggest a convenient solution.

2. DEVICE GEOMETRY.

The vertical geometry of the device to be discussed here is shown in Fig.1 and shows a simplified cross section of one half of one emitter, each of which is 100 microns wide. The heat source dimension is estimated from the doping concentrations of the materials and represents the volume associated with the collector depletion layer. Starting with this basic emitter structure many designs would be possible based on varying the emitter spacing or grouping a number of emitters to form cells from which larger power devices could be constructed.

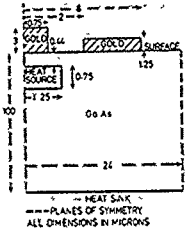


Fig.1. Cross section through the structure showing metallisation and dissipation region.

To simplify matters we will concentrate on the simple structure of Fig.1 and assume that the device has a large number of equally spaced emitter fingers, in this case 100um wide spaced by 48um. Use is made in the simulations of the planes of symmetry which exist in the structure

3. NUMERICAL METHODS AND A STATEMENT OF THE PROBLEM.

Three dimensional thermal simulation work has been undertaken using the finite element software ANSYS, and two in house simulators, one based on a finite difference technique and the other using the Transmission line matrix method (TLM). All these systems except the TLM method were capable of solving both steady state and transient thermal problems and take into consideration the variation of thermal conductivity with temperature. The TLM method proved to be very useful for simulating systems where lengthy time periods were involved and was used to obtain the curve shown in Fig.2.

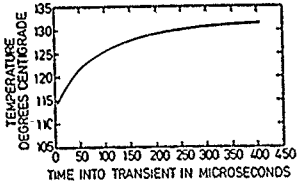


Fig.2. Peak Temperature at the end of repetitive power pulses. Duty Cycle=20% Pulse Length = 5 Microseconds.

Fig.2 shows the simulation result for a 20% duty cycle and 5 microsecond 'on' period. The temperature recorded is the peak value at the end of each power pulse. Notice that the peak temperature has not reached a maximum and the process is likely to take about 1

millisecond. A method which could obtain this result with less computation effort would be valuable and is the subject of the remainder of this paper.

4. AN ALTERNATIVE APPROACH TO THE PROBLEM.

Fig.3 shows the increase in the maximum temperature of the structure when a power pulse is applied. The peak temperature which would be reached is 324 degrees Centigrade neglecting any mounting layers (ie solder) which would typically raise the temperature a further 20 degrees. If the duty cycle of the device was very small this curve would accurately predict the maximum temperature for a required length of power input pulse.

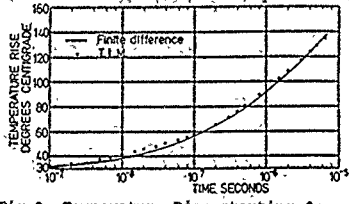


Fig.3. Temperature Rise starting from a Uniform Temperature of 30 degrees Centigrade.

It would be simple to produce such a curve for other duty cycles if the temperature distribution at the beginning of the power pulse was known. It seems that the only way to find this distribution would be to simulate the repetitive power cycle, the process we seek to avoid. The alternate approach is to first use a steady state simulator with an input power reduced by a percentage dictated by the duty cycle and use this distribution as an initial boundary condition for the start of the transient simulation. This power input is the average of that for each cycle. The temperature at some point in the bulk of the device will be oscillating about a mean value and the further away the point from the power source so this oscillation magnitude will decrease. In the case being studied an initial power pulse of 6.5 microseconds only produces a noticeable change in temperature within distances of about 30 microns of the source. At the point of maximum temperature the temperature simulated using this reduced input power will always be greater than that at the end of a repeating duty cycle, and therefore if this steady state distribution is used as an initial boundary condition for the commencement of a transient simulation, the resulting temperature will always be overestimated.

5. SOME NUMERICAL RESULTS.

The method described in the previous section was used to obtain the curves shown in Fig.4. They show the relationship between peak temperature, duty cycle, and length of input power pulse, for a given input power, base temperature and geometrical layout. Because of the simulation procedure used the results will always be pessimistic. To establish the accuracy of the curves a number of repetitive simulations were undertaken for a variety of

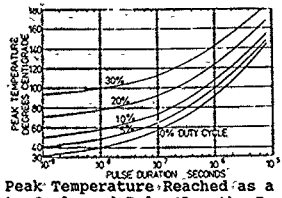


Fig.4. Peak Temperature Reached as a Function of Duty Cycle and Pulse Length. Base=30$_o$C

duty cycles and pulse lengths. Some typical results are shown in Fig.5 where the first second and fifth transient cycles are shown.

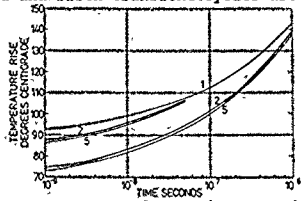


Fig.5. Temperature Rise during second, third and fifth cycles.

The temperature rises tested were always within 5% of the accurately simulated value, and the gains in computation effort were considerable.

6. DISCUSSION.

The method of using a steady state simulation and corresponding transient simulation has the potential to reduce the computation effort considerably. An estimate of the time saved to produce the curves shown in Fig.5, assuming 10 points/decade and that the transient simulation will give a reasonably accurate peak temperature within 10 cycles, is a factor of about 500. A 5% error in the predicted temperature rise is in reality quite satisfactory as problems of reproducibility in bonding layers can be comparatively significant. The results of the simulation using all the simulators indicated in section 3 gave very similar results. The TLM method was by far the most useful in the simulation of transients involving large time periods, but it was not possible to couple it to a steady state simulator and use the method developed here. Also the TLM method requires relatively large amounts of data space, about 8 times the equivalent finite difference approach.

7. CONCLUSION.

The method outlined in this paper shows that the time required to compare and assess the thermal design of proposed HBJT structures for use in transient applications may be reduced considerably with a modest compromise in the accuracy obtained.

8. ACKNOWLEDGEMENT.

The authors acknowledge the help of Plessey Research Caswell in supplying typical geometrical and process design data.

2D NUMERICAL MODELLING OF AlGaAs/GaAs HBTs

NEIL P. WALDIE, PHILIP A. MAWBY and ANDREW McCOWEN

Department of Electrical & Electronic Engineering
University College of Swansea,
SWANSEA, UK.

Abstract - A general purpose 2D device simulator (SUDS) has been enhanced to investigate the specific mechanisms which cause DC gain degradation in AlGaAs/GaAs heterojunction bipolar transistors (HBTs). Access to the distributions of potential and electron and hole concentrations from the numerical simulator greatly aid the analysis which shows that surface recombination as opposed to surface Fermi level pinning is responsible for the degradation of gain.

I INTRODUCTION

AlGaAs/GaAs heterojunction bipolar transistors are becoming increasingly important for applications in microwave integrated circuits due to their very high switching speeds. In the fabrication of these devices either an etch down to the AlGaAs layer or an implant is used to isolate the emitter and base contact regions. It is generally understood that surface effects, namely surface Fermi level pinning and surface recombination are the causes of gain degradation in both of these structures [1].

The modelling of such highly non-linear two-dimensional effects is best achieved by a numerical device simulator which solves the basic equations of Poisson and current continuity. A general purpose device simulator (SUDS), capable of handling arbitrary doping and geometry, has been developed and recently enhanced to handle heterojunctions and surface effects. In this paper we have studied the effects of different configurations surface charge and recombination centres and have isolated the mechanisms which significantly control the basic current causing the degradation in gain.

II THE NUMERICAL MODEL

The device modelling package used in this work solves the semiconductor heterojunction device equations on a arbitrarily shaped two dimensional domain. The physics of the semiconductor are modelled by a set three partial differential equations, namely, Poisson's equation

$$\nabla \cdot \underline{D} = \rho \qquad (1)$$

and two continuity equations one for the electron distribution and one for the hole distribution.

$$\nabla \cdot \underline{J}n = qU \qquad (2)$$

$$\nabla \cdot \underline{J}p = -qU \qquad (3)$$

These fundamental equations require constitutive relationships for the vector flux terms. The electric displacement term $\underline{D}$, is expressed as

$$\underline{D} = \epsilon \underline{E} \qquad (4)$$

and the current densities are expressed in terms of the modified drift-diffusion relationships

$$\underline{J}n = -q\mu_n n \nabla (\psi + \phi_n) + qD_n \nabla n \qquad (5)$$

$$\underline{J}p = -q\mu_p p \nabla (\psi + \phi_p) - qD_p \nabla p \qquad (6)$$

where the φ_n and φ_p parameters described the compositional variation of the semiconductor [2]. The basic set of equations (1) - (3) are discretised using the control region method [3], on a triangular mesh.

The method involves the integration of each of the equations over the control (Voronoi) area associated with each mesh node. By application of the divergence theorem the resulting surface integrals can be converted to line integrals around the perimeter of the control area. So that for a node i in the mesh the discretised equations become

$$F_\psi(\psi) = \sum_{j=1}^{M_i} D_{ij} d_{ij} - \rho_i A_i = 0 \qquad (7)$$

$$F_{\varphi_n}(\varphi_n) = \sum_{j=1}^{M_i} J_{nij} d_{ij} - qU_i A_i = 0 \qquad (8)$$

$$F_{\varphi_p}(\varphi_p) = \sum_{j=1}^{M_i} J_{pij} d_{ij} + qU_i A_i = 0 \qquad (9)$$

where M_i is the number of nodes connected to node i. The constitutive relations (4) - (6) are then discretised using the standard finite difference method in the case of $\underline{D}$ the electric field displacement and by the Scharfetter-Gummel [4] method in the case of electron and hole currents along edge ij. The resulting set of non-linear equations ($F_\psi(\psi)$, $F_{\varphi_n}(\varphi_n)$ $F_{\varphi_n}(\varphi_n)$), are solved using the Newton-Raphson method for the solution variables ψ, φ_n and φ_p, using either a Gummel decoupled scheme or a coupled scheme.

The large linear equation set formed at each Newton iteration is solved using the ICCG method, for symmetric matrices and the CGS method for non-symmetric. Both of which use an extremely compact sparse storage scheme which minimises the memory requirements.

III THE PHYSICAL MODELS

The physical models described here are for the AlGaAs/GaAs material system, however, SUDS has built-in models for most other common semiconductors. The mathematical models describing physical processes require a number of empirical parameters to fit them to the specific material.

A. Mobility

Since the AlGaAs/GaAs material system is a ternary compound, the mobility model must take account of the extra scattering caused by the alloy, in addition to the standard phonon and impurity scattering mechanisms of the homogeneous material. Experimentally this processes manifests itself as a reduction in the low-field mobility which decreases as alloy mole fraction (x) increases. The mobility model implemented in the current work is thus a function of temperature, net doping, alloy composition and electric field [5][6]. The standard expression is used for lattice scattering

$$\mu_L = \mu_o \left(\frac{T}{300}\right)^{-\alpha} \qquad (10)$$

where T is the lattice temperature, μ_o is the low field mobility in GaAs and α is an experimental fitting term. Impurity scattering is modelled by

$$\mu_{LI} = \mu_L + \frac{\mu}{1 + \left(\frac{\Gamma}{N_{ref}}\right)^{\alpha}} \qquad (11)$$

where Γ is the net doping and the fitting parameters μ, N_{ref} and α are determined from experimental data [5].

The high field mobility of carriers in small geometry devices is of critical importance. The model used here for electrons is the well known model for III-V materials [6], which includes the intervally scattering of electrons causing negative differential resistance in lightly doped structures.

$$\mu_n(E) = \frac{\mu_{LI} + \frac{V_s}{E}\left(\frac{E}{E_c}\right)^4}{1+\left(\frac{E}{E_c}\right)^4} \qquad (12)$$

where V_s is the saturation velocity and E_c is the critical field. Velocity overshoot effects are not included here since they are localised to the collector depletion region, and so play no role in the phenomena currently under consideration.

B. Recombination

Bulk recombination is modelled by the Schockley-Hall-Read (SHR) process and are characterised by two minority carrier lifetimes and

$$R = \frac{pn - n_i^2}{\tau_n(p+n_i)+\tau_p(n+n_i)} \qquad (13)$$

The lifetimes are dependent on the local net doping. Auger and Impact Ionisation are also options but are not selected for this device.

Surface effects are modelled by the incorporation of surface charge and a separate surface recombination mechanism. Surface charge is defined as a fixed donor or acceptor charge, which is an additive term to ρ, in Poisson's equation (1). The increased recombination at the surface is assumed to arise from discrete energy levels which can be defined to be donor of acceptor levels. For the trap at each energy level a separate SHR model is established.

$$R_s = \frac{n_s p_s - n_i^2}{\tau_{ns}(p_s+p_T) + \tau_{ps}(n_s+n_T)} \qquad (14)$$

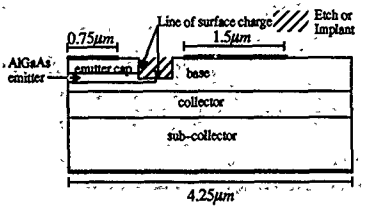
where τ_{ns} and τ_{ps} are the minority carrier lifetimes associated with the trap, and n_T and p_T are the electron and hole concentrations that would exist in equilibrium if the Fermi level sat at the energy of the trapping centre.

IV RESULTS

Fig. 1 shows a typical AlGaAs/GaAs HBT used in microwave integrated circuits. Fig.2 shows a comparison of the IV characteristics of the structure with and without surface effects, and clearly demonstrates the significant increase in the base current below approximately $V_{BE} = 1{\cdot}1V$. Fig. 2 also shows the position of the traps and the pn base-emitter junction in the AlGaAs layer. The surface charge was set to $2{\cdot}6 \times 10^{11} cm^{-2}$ and the trap depth was set to the intrinsic Fermi level E_{Fi}. With surface recombination switched out of the simulator, the base current reverted back to its former level demonstratung that the surface charge alone had little influence on the terminal base current.

V CONCLUSIONS

The highly non linear models associated with surface effects have been incorporated into SUDS to demonstrate their effects on the characteristics of the AlGaAs/GaAs HBT. Surface recombination is shown to be the primary cause of gain degradation at low bias levels.



emitter cap	0.15mm	n^+	GaAs	$5 \times 10^{17} cm^{-3}$
emitter	0.12mm	N	AlGaAs	$5 \times 10^{17} cm^{-3}$
base	0.13mm	p^+	GaAs	$4 \times 10^{18} cm^{-3}$
collector	0.35mm	n	GaAs	$1 \times 10^{16} cm^{-3}$
sub-collector	0.7mm	n^+	GaAs	$4 \times 10^{18} cm^{-3}$

Fig.1. Typical structure of AlGaAs HBT

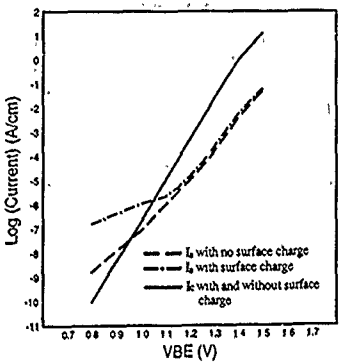


Fig.2. I-V characteristics for structure in Fig.1.

REFERENCES

1. C.H.HENRY, R.A.LOGAN and F.R.MERRIT, "The effects of surface recombination on current in $Al_xGa_{1-x}As$ heterojunctions", *J. App.Phys* 49(6) June 1978.
2. J.E.SUTHERLAND, J.R.HAUSER, "A computer analysis of heterojunctions and graded composition solar cells" *IEEE Trans.* on Electron Devices, Vol ED-34, No 4, April 1977.
3. B.J.McCARTIN, R.H.HOBSS, R E.LaBARRE, P,E.KIRSCHNER and G.A.PETERSON,"Discretization of the Semiconductor Device Equations", In: Proc. of short course held at NASCODE IV, conf. Ed. J.J.Miller Boole Press p72, 1985.
4. D L SCHARFETTER and H.K.GUMMEL, "Large Signal Analysis of Read Diode Oscillators", IEEE Trans. Electron Devices, Vol. IEEE-ED 16, pp 64-77, 1969.
5. C. AMANO, H. SUGIURA, M.YAMAGUCHI and K.HANE, "Fabrication and analysis of AlGaAs/GaAs tandem solar cells with tunnel interconnections", *IEEE Trans.* on Electron Devices, Vol 36,No 6,June 1989.
6. J.YOSHIDA, M. KURATA, K.MORIZUKA and A.HOJO, "Emitter-base bandgap grading effects on AlGaAs/GaAs heterojunction bipolar transistor characteristics", *IEEE Trans.* on Electron Devices, Vol.32,No.9 Sept 1985.

PRACTICAL APPLICATION OF NUMERICAL MODELLING TO HETEROJUNCTION BIPOLAR TRANSISTOR DESIGN

Robert E Miles, David J Holder and Christopher M Snowden
Department of Electronic and Electrical Engineering, University of Leeds,
Leeds, England, LS2 9JT

Abstract

A two-dimensional simulation of the heterojunction bipolar transistor is described which has been used to understand the processes taking place in real devices. In this paper the effects of electron/hole recombination in implanted regions is discussed and it is also shown that misalignment between the doping and alloy junctions at the emitter/base interface can significantly degrade the device performance.

INTRODUCTION

Mathematical models have long been used to understand and hence improve the operation of semiconductor devices. These models, at their simplest, can give a closed from analytical solution which in the past has been instrumental in the development of improved structures. However for modern devices, with their reduced dimensions and high internal electric fields, the approximations necessary for an analytical solution are no longer valid. It has therefore been necessary to develop numerical models where more accurate descriptions of the device physics can be included albeit at the expense of solution time.

This paper will describe how a two-dimensional model has been used to understand and aid the design of heterojunction bipolar transistors (HBTs).

THE MODEL

The device modelled in this work is the GaAs/AlGaAs HBT where the wide band gap emitter is the AlGaAs alloy with the mole fraction of aluminium in the region of 0.2 to 0.3 and the base is GaAs. The emitter/base junction can be abrupt or graded. Being a bipolar device the recombination of electrons and holes in an HBT is an important factor which in this model is described by the Shockley-Hall-Read mechanism via recombination states at mid gap. The model allows for a wide variation of recombination rates in different regions of the device to account for the effects of surfaces, material interfaces and various implants. Figure 1 shows the basic geometry of the devices modelled. In practice, owing to the symmetry of the structure only one half of the device needs to be simulated with a consequent saving in computer run time.

A drift-diffusion description of the HBT has been implemented taking into account the dependence of carrier velocity on electric field and the variation of the semiconductor properties across the junction. The device equations are solved by finite difference techniques on a two-dimensional grid which is automatically refined to a smaller mesh size in regions where the electrical potential is varying rapidly.

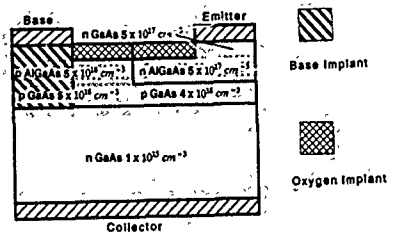


Fig. 1. Heterojunction Bipolar Transistor structure modelled in this work. (NB ½ of device shown)

A number of effects can be studied with the aid of this model but this paper will concentrate on two areas where it has been used to interpret the measured performance of actual devices fabricated at the GEC/Plessey Research Centre, Caswell. The two areas are firstly, the effect of an implanted oxygen isolation layer between the emitter and base and secondly, the consequences of a misalignment between the doping and alloying junctions at the base emitter interface. The misalignment problem has also been studied using a one dimensional model of a p-n heterojunction because it is much faster to run than the full two dimensional case.

THE EMITTER/BASE ISOLATION IMPLANT

The HBT shown in Figure 1 is a planar structure which has a number of advantages during fabrication. A less desirable feature however is that the base is in contact with the emitter across a vertical plane extending down from the surface. This effectively means that electrons will be injected into a region where the base is much wider than the optimum (as defined by the region under the emitter). The injected electrons will therefore take a longer time to traverse the base and there will also be a higher probability of their recombining on the way. These effects will lower both the frequency response and the current gain of the device - both undesirable effects. One solution to this problem is to introduce the isolation layer shown in Figure 1 between the emitter and base by means of an oxygen implant which effectively creates high resistivity material in the required region.

The oxygen implant effectively creates a region of heavy crystalline damage which means that there will be a high density of localised energy levels in the forbidden gap which will in turn make the electron/hole recombination process much faster This implant is therefore modelled as a region having a much shorter recombination time than the rest of the device.

Figure 2 shows the effect of this recombination on the current gain. In this Figure, device A has a recombination time of 5×10^{-9}s in the GaAs base and 2×10^{-10}s in the implanted AlGaAs base contact. Device B is the same but with the addition of a recombination time of 4×10^{-11}s in the isolation implant. The current gain of about 30 for device B is typical of what is observed in practice. The model was also used to investigate the effect of recombination at the top surface and the emitter/base interface. It was found that recombination at the top surface had little or no effect while a recombination centre density of greater than $10^{13}cm^{-3}$ at the emitter/base interface was needed to account for the observed reduction in current gain. This density is much higher than would be expected for modern technology.

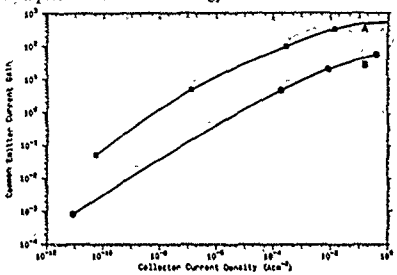


Fig. 2. The effect of recombination in the isolation region on the current gain.

By using this model it was therefore shown that out of a number of possible causes, excessive recombination in the isolation implant region was the most likely reason for the poorer than expected performance of these devices. It would have been difficult to isolate the effects of recombination in the different regions without the aid of a suitable model.

ALLOY/DOPING MISALIGNMENT

Simple heterojunction theory assumes that the boundary between the n and p regions corresponds exactly with the change from AlGaAs to GaAs. In practice, owing to the limitations of the technology a misalignment of these junctions can easily occur. In the devices studied here the most likely reason for this misalignment is the diffusion of acceptors from the highly doped GaAs base into the AlGaAs emitter. This effect is illustrated in Figure 3 which shows the movement of the p-type doping (shaded black) into the n-type emitter (hatched) which, for the purposes of this paper, is defined as a positive shift.

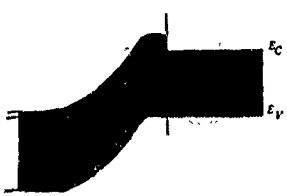


Fig. 3. Schematic representation of a misalignment between the doping and alloy junctions.

A one dimensional p-n heterojunction model has been used to study the effect of misalignment on the emitter injection efficiency. The emitter injection efficiency, defined as the percentage of the total emitter/base current carried by the emitter majority carriers (in this case electrons), is the most important quantity affecting the gain of the device. The valence band discontinuity in an HBT virtually eliminates the hole current giving an efficiency of close to 100% and hence a very high gain. Figure 4 shows the reduction in emitter injection efficiency caused by various positive shifts up to 40 nm. (Negative shifts give little or no effect.) Shifts of a few nanometres, corresponding to movements of the junction through about 10 atomic layers, give significant reductions in the injection efficiency, sufficient to have a significant effect on the current gain of an HBT. (factors of 10 are typical.)

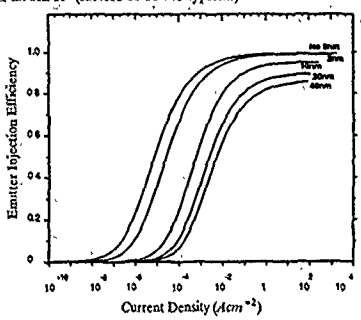


Fig. 4. The effect of misalignment on the emitter injection efficiency.

The reduction in the injection efficiency levels off for shifts greater than about 40 nm because at this point, for the doping levels considered, the depletion regions associated with the p-n junction are all within the emitter and the alloy junction plays no further part in the transistor action. It is also expected that graded alloy junctions will be less susceptible to the effects of misalignment.

CONCLUSION

This paper has described just two technological problems that can have a bearing on the performance of an HBT. By using the model individual features can be studied in isolation, a process which is difficult in real devices because of the uncertainties of the fabrication process and characterisation techniques. This in turn can be fed back to the device engineer to give insight into the important parameters which must be controlled during fabrication. Without such a model, it is very difficult to identify these parameters.

MONTE CARLO SIMULATIONS OF TRANSPORT IN P-TYPE HETEROSTRUCTURES

R. W. Kelsall and R. A. Abram

Applied Physics Group, School of Engineering and Applied Science,
University of Durham, South Road, Durham, DH1 3LE, UK.

Abstract

A Monte Carlo transport simulation is described which incorporates a realistic description of the valence bandstructure in semiconductor quantum wells, including heavy – light hole mixing and strain effects. The simulation is used to analyse hole transport in lattice matched GaAs/AlAs and pseudomorphic InGaAs/GaAs structures. The 77K phonon-limited mobility in the pseudomorphic well is nearly six times that in the lattice matched well, but much of this enhancement is removed by alloy scattering.

Introduction

For several years there have been aspirations to develop a complementary logic process technology in III–V semiconductors. Such a logic system ought, in principle, to combine the extremely low stand-by power dissipation of CMOS with the high speed of the III–V devices. However, no practical system has yet been realised, a major setback being the limited speed achieved in the p-type devices. The incorporation of strained InGaAs layers in modulation doped FETs has opened up a new route to a potentially fast p-type device. It is found that the axial strain present in a pseudomorphic heterostructure shifts the energy levels of the heavy hole subbands relative to the light hole subbands. This can lead to a reduced effective mass in the highest valence subband, which would be expected to give rise to enhanced hole mobilities and saturation drift velocities [1]. However, the mobilities and drift velocities measured to date in p-type pseudomorphic InGaAs heterostructures have been disappointingly low [2–4], showing little improvement over the values obtained in lattice matched systems. The reasons for this discrepancy between expected and observed results have not been identified, and no theoretical analysis has been undertaken.

Any such analysis of p-type heterostructures must take into account the detailed form of the valence subband structure, which can be severely distorted from the usual near-parabolic form by quantum-size-induced mixing of the heavy and light hole states. This mixing also affects the hole wavefunctions, which are important in determining the transition rates for the various scattering processes active in the device. We have developed a Monte Carlo simulation scheme which satisfies all the above requirements. The hole subband energy dispersions and wavefunctions are calculated using the k.p method, and are then used to calculate the scattering rates as a function of hole wavevector. Hole transport is then simulated in the plane of the heterostructure active layers taking into account the subband energies and group velocities, and the angular dependences of the heterostructure scattering rates. We have investigated two cases: i) a lattice matched GaAs/AlAs single quantum well and ii) a pseudomorphic InGaAs/GaAs single quantum well. In each case we have taken the lattice temperature to be 77K, in order to facilitate comparison with a range of experimental results. These simulations represent, to the best of our knowledge, the first detailed models of hole transport in p-type heterostructures, allowing us to investigate the origins of the speed limitations of both lattice matched and pseudomorphic devices.

Calculations

Figures 1a and b show the in-plane energy dispersion of valence subbands in the GaAs lattice-matched and InGaAs pseudomorphic quantum wells respectively. The GaAs quantum well width is 100Å, whilst the InGaAs well width is 90Å and the In concentration is $x = 0.18$. The bandstructures show relatively little anisotropy in the quantum well plane, and we have assumed them to be isotropic in the following calculations. Strong heavy – light hole mixing is clearly evident in the GaAs well, with severe distortion of the subbands near regions of repulsion ('avoided crossings') between adjacent subbands. The second and fourth highest subbands have energy minima which are displaced from the zone centre, leading to regions of extremely high densities of states. The effective mass at the zone centre of the highest subband is 0.156, but rises to a maximum value of 0.901 due to interaction with the next highest subband. By comparison the valence bandstructure in the InGaAs well is much less affected by mixing, and the effective mass in the highest subband varies from 0.126 at the zone centre to 0.434 for an in-plane wavevector k_t of 0.1Å^{-1}.

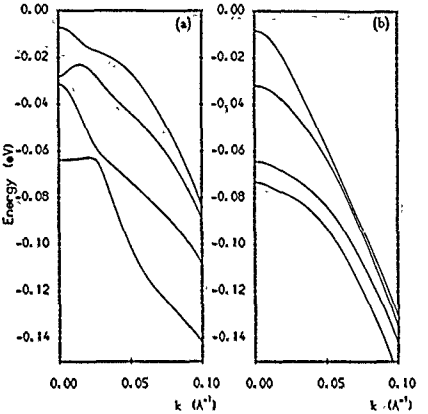


Figure 1 In-plane valence subband dispersion for (a) a GaAs/AlAs lattice matched and (b) an InGaAs/GaAs pseudomorphic quantum well.

The hole-phonon scattering rates are dependent on the subband energies, the quantum-confined hole wavefunctions and the densities of states. We have obtained this information from the k.p calculations for each required value of k_t, which enabled us to calculate the integrated hole-phonon scattering rates out of any initial k_t state into any allowed final state, via intra- and inter-subband transitions. We have included optical phonon scattering via the Fröhlich (polar) and deformation potential interactions, and acoustic deformation potential scattering. Rates for piezolectric scattering were calculated for the GaAs quantum well, but found to be negligible, and the calculations were not repeated for the InGaAs case.

In figure 2 we have shown the total hole-phonon scattering rate in the highest valence subband of the GaAs and InGaAs quantum wells *vs.* initial energy. The features labelled on the diagram correspond to thresholds for the principal transitions. Polar optical scattering is the dominant process in both quantum wells; however, all processes are considerably stronger in the GaAs well. This is due to the enhanced density of states in the GaAs valence subbands, the spike-like structures at the thresholds for 1–2 and 1–4 scattering correspond to the peaks in the density of states at the off-zone-centre energy minima in the second and fourth subbands.

For the InGaAs quantum well, alloy scattering must also be considered. We calculated rates for intrasubband alloy scattering using the 2D density of states, but taking the bulk form of

the scattering matrix element [5]. We are currently working on calculations of the matrix element using the quantum confined k.p wavefunctions in the manner described above for phonon scattering. There appears to be no information available on the alloy scattering potential ΔE for holes in InGaAs. One possible prescription is to take the difference between the band edge energies of the constituent binary compounds, measured relative to the vacuum level. This gives a value of $\Delta E = 0.267$eV. In our simulations we have also used the value of $\Delta E = 0.42$eV determined empirically for electrons in InGaAs [6].

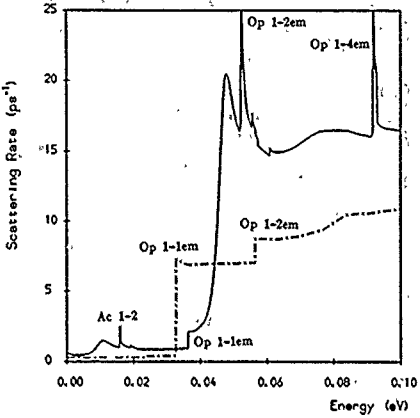


Figure 2 Total 77K hole-phonon scattering rate in the highest valence subband: — GaAs/AlAs, · – · InGaAs/GaAs quantum wells. The labels mark the thresholds for: Op 1–1em, 1–2em, 1–4em; optical phonon emission scattering into the highest, second highest and fourth highest subbands respectively, and Ac 1–2, acoustic phonon scattering into the second subband.

Results and Discussion

Figure 3 shows the 77K velocity-field characteristics for GaAs lattice matched and InGaAs/GaAs pseudomorphic heterostructures. For the InGaAs system we have shown results calculated without alloy scattering, and with alloy scattering potentials of 0.267 and 0.42eV. The phonon-limited drift velocities in the InGaAs quantum well are, as expected, considerably larger than the measured and simulated values in GaAs heterostructures. The simulated phonon-limited low field mobility is approximately 38,000$cm^2V^{-1}s^{-1}$, which is nearly six times larger than our simulated value of 6,400$cm^2V^{-1}s^{-1}$ for the GaAs well. Alloy scattering gives rise to significant reductions in the drift velocity and mobility; we obtained mobilities of approximately 15,000 and 10,000$cm^2V^{-1}s^{-1}$ with alloy scattering potentials of 0.267 and 0.42eV respectively. Even this latter result is some 33% larger than the highest experimental results obtained [2], and the drift velocities at higher fields are well above the experimental values shown in figure 3. In contrast, the simulated characteristic for the GaAs quantum well agrees well with the two experimental curves shown, both measured on single GaAs heterojunctions. The simulated low field mobility also agrees well with experimental data [8,9].

The low mobilities observed experimentally in the InGaAs system could be due to impurity scattering, which was not included in our simulations. However, there is no clear reason why this process should be any more important in the InGaAs than in the GaAs quantum well. One process which may well be more prevalent in the InGaAs system is interface roughness scattering, since the InGaAs/GaAs growth technology is not as well established as that for GaAs/AlGaAs and GaAs/AlAs interfaces. More work on alloy scattering is also needed to determine whether a higher upper limit than that used here is appropriate.

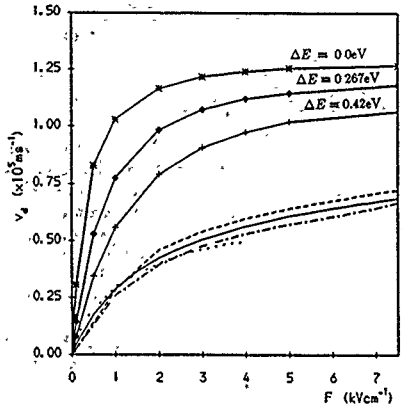


Figure 3 77K hole drift velocity *vs.* in-plane electric field. — GaAs simulation; ∘, ◇, +, InGaAs simulations; · – ·, · · · GaAs experiment [7]; · · · InGaAs experiment [4].

Conclusions

The Monte Carlo-k.p model described above represents a valuable tool for analysing hole transport in p-type heterostructures. Our simulations predict that the 77K phonon-limited drift velocities and low field mobility in a pseudomorphic InGaAs/GaAs quantum well are significantly larger than in a lattice matched GaAs/AlAs well. Alloy scattering is important in reducing the mobility in the InGaAs system, but we suspect that other, hitherto unidentified scattering processes are also responsible for the low values obtained experimentally.

We wish to acknowledge E. P O'Reilly, W. Batty and A. C. G. Wood for their helpful advice and the provision of band-structure data.

References

[1] O'Reilly E P *Semicond. Sci. Technol.* 4 121 (1989)
[2] Fritz I J, Drummond T J, Osbourn G C, Schirber J E and Jones E D *Appl. Phys. Lett.* 48 1678 (1986)
[3] Lancefield D, Batty W, Crookes C G, O'Reilly E P, Adams A R, Homewood K P, Sundaram G, Nicholas R J, Emeny M and Whitehouse C R *Surface Science* 229 122 (1990)
[4] Reddy M, Grey R, Claxton P A, and Woodhead J *Semicond Sci. Technol.* 5 628 (1990)
[5] Harrison J W and Hauser J R *Phys. Rev.* B 13 (1976)
[6] Marsh J H, Houston P A and Robson P N *GaAs and Related Compounds 1980* IOP Conference Series No. 56 p621 (1981)
[7] Masselink W T, Braslau N, LaTulipe D, Wang W I and Wright S L *GaAs and Related Compounds 1987* IOP Conference Series No. 91 p665 (1988)
[8] Störmer H L, Gossard A C, Wiegmann W, Blondel R and Baldwin K *Appl. Phys. Lett.* 44 139 (1984)
[9] Mendez E E and Wang W I *Appl. Phys. Lett.* 46 1159 (1985)

The Numerical Fourier Method for Multidimensional Semiconductor Device Modeling

V. Axelrad*, S. Eckart
Chair for Integrated Circuits, Technical University Munich, Germany

Introduction

High numerical accuracy is essential to achieve convergence and reliable results for numerical models of advanced semiconductor devices. Physical effects such as impact ionization and local carrier heating sensitively depend upon the carrier current density, which can be of insufficient numerical accuracy in classical numerical models. While these accuracy problems can limit the quality of the results, possible convergence problems are of even more practical significance.

A way to improved accuracy of the solution is the use of higher-order methods. Spectral methods offering infinite-order accuracy $O(e^{-N})$ (N is the number of degrees of freedom) are particularly attractive. Recently developed algorithms demonstrated the applicability of a spectral Fourier-series discretization to strongly nonlinear semiconductor device problems [1, 2]. Limitations of the spectral approach with respect to possible geometries and boundary conditions have been addressed and efficient numerical algorithms for the solution of the spectral equations developed.

Localized Sampling

In order to achieve high accuracy, locally fast variation of the state variables has to be sufficiently resolved by the underlying discretization mesh. Reasonable solution time can, however, only be maintained by local refinement of the mesh. A general approach is the use of coordinate transformations and equidistant mesh spacing in the new coordinate system. A special case is a separable (tensor-product type) coordinate transformation, in 2D:

$$\nabla_{x,y} f = \begin{pmatrix} \frac{d\xi}{dx} & 0 \\ 0 & \frac{d\eta}{dy} \end{pmatrix} \cdot \nabla_{\xi,\eta} f$$

Spectral accuracy is assured by using infinitely differentiable mapping functions $\xi, \eta \in C^\infty$ (i.e. by subjecting their series representations to exponential convergence requirements).

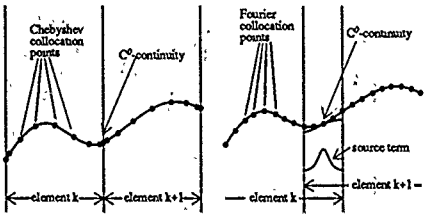


Figure 1 Spectral element discretization based on Chebyshev polynomials (left) and Fourier (cosine) series (right)

To alleviate the restrictions imposed by the rectangular tensor-product mesh, spectral element discretization is under consideration. The standard method is to use nonequidistant Chebyshev polynomials (Fig. 1, left). The associated matrices are, however, not optimal for a numerical solution. The space domain matrix exhibits an algebraic singularity at the boundaries [3], whereas the spectral domain matrix is ill-conditioned [4].

An approach using overlapping high order elements using cosine basis functions and coupling the elements through source terms controlled by the function difference between both elements (Fig. 1, right) appears to be promising. The associated spatial domain matrices possess advantageous properties. Preliminary one-dimensional results show the feasibility of this approach.

Solving the Equations

The algorithm used for the solution of the discrete equations is of central importance for the cpu-time and memory requirements of the method. In general, direct methods (Gaussian elimination) are not feasible, since the system matrix of the spectral equations is not sparse.

*presently at. Technology Modeling Associates, Palo Alto, California, USA

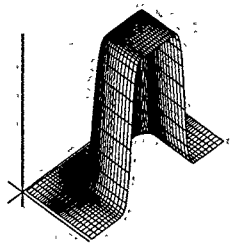


Figure 2: Net doping concentration in the JFET.

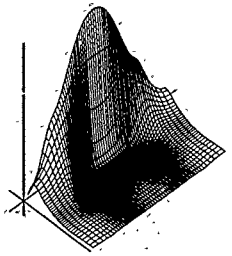


Figure 4: Hole current density.

Iterative techniques such as incomplete LU-decomposition can be applied to the spectral-domain equations taking advantage of their diagonal dominance and well-conditioning [2]. Alternatively, iterative algorithms based on space-domain preconditioning of the spectral equations by a finite-difference operator [5] can be used. This hybrid spatially preconditioned spectral solution is attractive especially for complex multidimensional device models.

Application

An application example is provided by a p-channel JFET, $L_{Gate} = 1.5\mu$m. A drain-voltage of -0.1V and a gate-voltage of $+0.1$V is applied. Fig. 2 shows the doping profile (p-area: 1×10^{15}cm^{-3}, n-area: 3×10^{15}cm^{-3}), Fig. 3 shows the space charge density and Fig. 4 the hole current density ($j_{max} = 3.3$A/cm^{-2}). Contacts were defined in the two corners of the device $(0,0)$ and $(0,L_y)$ and at $(L_x, L_y/2)$ (i.e. source, drain and gate respectively) using the distributed voltage-controlled current source model [2]. The simulation was performed on a 32×64 point non-equidistant grid covering an area of 2μm $\times$ 4μm.

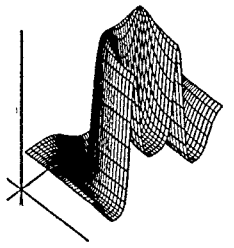


Figure 3: Space charge density.

Summary

Basic functionality of the method has been demonstrated in previous work. Current efforts address localized meshing via high-order element patching, effective iterative solution of the spectral equations and refinement of the physical models (energy balance equations) taking advantage of the method's high numerical accuracy. Comparison of the method's properties to those of classical low-order space-domain methods is being carried out.

References

[1] V. Axelrad. Fourier approach to semiconductor device modelling. *Int. Journal of Numerical Modelling*, 2:31–52, 1989.

[2] V. Axelrad. Fourier method modeling of semiconductor devices. *IEEE Trans. on CAD*, 9(11), 1990.

[3] B. Fornberg. An improved pseudospectral method for initial-boundary value problems. *J. Comput. Phys.*, 91:381–397, 1990.

[4] D. Gottlieb and S.A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*. CBMS-NSF Monograph No. 26. SIAM, Philadelphia, 1977.

[5] V. Axelrad. A space-domain preconditioned spectral method for nonlinear boundary value problems. *submitted*.

GENERALISATIONS OF THE BOX METHOD WITH APPLICATIONS TO THE SEMICONDUCTOR PROBLEM

W H A. Schilders
Philips Research Laboratories
Applied Mathematics Group
Building WAY, Room 2 09
PO Box 218, 5600 MD, Eindhoven (NL)

ABSTRACT

In this paper we present the box method as a mixed finite element method with a suitable quadrature. This interpretation provides a way of defining a continuous current density for the box method. Also, in this way, higher order box methods can be derived which are suitable for application to the discretisation of the semiconductor device problem.

1. INTRODUCTION

The box method is the most widely used discretisation method for the semiconductor device problem. The main reason for this is that it yields discrete current conservation and utilizes 'upwinding' by using the Scharfetter-Gummel expressions for the current densities. Another reason for its use is that solutions of the resulting discrete systems satisfy the maximum principle (positive carrier concentrations !) if the underlying triangular mesh is of Delauney-type. A drawback of the method is that it does not yield a continuous expression for the electric field and the current densities, only components along the edges of the meshes are given. Because of the latter, one could be tempted to think about the application of finite element methods to the discretisation of the semiconductor problem. Unfortunately, it is wellknown that ordinary finite element methods, although frequently used in other disciplines, do not possess any of the properties listed above (although 'upwinding' can be achieved, for example, by using the streamline upwind methods proposed by Hughes and Brooks, cf. [1]).

Recently, mixed finite element methods have been proposed for the discretisation of the semiconductor device problem (cf. [2]). These methods differ from ordinary finite element methods in several respects but, most importantly, they yield Scharfetter-Gummel type expressions for the current densities in a natural way and resulting solutions satisfy discrete current conservation. Unfortunately, the mixed finite element methods proposed have been shown to be rather unstable (cf. [3]). For example, the solution of Poisson's equation will lead to unphysical oscillations in the electric potential ψ, even for triangulations without any obtuse angle. The same holds for the solutions of the continuity equations whenever a non-zero recombination term is present. Furthermore, if obtuse angles occur, the method will always be unstable, even when the underlying mesh is of Delauney-type. Although attempts have been undertaken to remedy this situation (cf. [3] where quadrature rules are suggested, and [4] where new elements are introduced), the latter problem has not been resolved so far.

In the following sections we present a class of mixed finite element methods which can be used on Delauney-type triangular meshes. The box method is shown to be equivalent to the lowest-order element in this class. Because of this, a recipe can be given for producing a continuous expression for the electric field and the current densities from the solutions obtained with the box discretisation. Furthermore, higher order extensions of the box method are immediate consequences of this new class of mixed finite element methods.

2. MIXED FINITE ELEMENT FORMULATION OF THE SEMICONDUCTOR PROBLEM

We consider the following system of equations.

$$\nabla \cdot \mathbf{J} = R(u), \tag{1}$$

$$\mathbf{J} = a\nabla u. \tag{2}$$

on the region $\Omega \subset R^2$, with suitable boundary conditions. Remark that Poisson's equation as well as the continuity equations for holes and electrons can be written in the above form.

The mixed variational formulation of the problem (1)-(2) is: find $(u, \mathbf{J}) \in L^2(\Omega) \times H(div, \Omega)$ such that

$$\int\int_\Omega \phi \nabla \cdot \mathbf{J} d\Omega = \int\int_\Omega \phi R(u) d\Omega \qquad \forall \phi \in L^2(\Omega) \tag{3}$$

$$\int\int_\Omega a^{-1}\mathbf{J} \cdot \mathbf{r} d\Omega = -\int\int_\Omega u \nabla \cdot \mathbf{r} d\Omega \qquad \forall \mathbf{r} \in H(div;\Omega) \tag{4}$$

where

$$H(div;\Omega) = \{\mathbf{r} \in (L^2(\Omega))^2 | \nabla \cdot \mathbf{r} \in L^2(\Omega)\}$$

It is wellknown that the problem (3)-(4) has a unique solution $(u, \mathbf{J}) \in L^2(\Omega) \times H(div, \Omega)$, which is the weak solution of (1)-(2).

The mixed discretisation now consists of choosing suitable finite dimensional subspaces $V_h \subset L^2(\Omega)$ and $W_h \subset H(div, \Omega)$ and to restrict the problem (3)-(4) to these subspaces: find $(u_h, \mathbf{J}_h) \in V_h \times W_h$ such that

$$\int\int_\Omega \phi_h \nabla \cdot \mathbf{J}_h d\Omega = \int\int_\Omega \phi_h R(u_h) d\Omega \qquad \forall \phi_h \in V_h \tag{5}$$

$$\int\int_\Omega a^{-1}\mathbf{J}_h \cdot \mathbf{r}_h d\Omega = -\int\int_\Omega u_h \nabla \cdot \mathbf{r}_h d\Omega \qquad \forall \mathbf{r}_h \in W_h \tag{6}$$

The remaining problem is to construct suitable subspaces V_h and W_h. The Brezzi-Babuska conditions (cf. [5]) give sufficient (but not necessary) conditions on these subspaces in order to guarantee uniqueness of the solution of (5)-(6). For triangular meshes, a class of finite element spaces have been proposed which satisfy these conditions, the so-called *Raviart-Thomas elements* (cf. [6]). The latter are widely used in the context of mixed finite element discretisations.

3. EQUIVALENCE OF THE BOX METHOD AND A LOW ORDER MIXED FEM

In view of the foregoing we now assume that the region Ω has been triangulated and that the mesh is of Delauney-type. Because of the latter assumption we can associate, with each mesh point, a so-called *box* which is constructed by intersecting the midperpendiculars of the edges of the triangles. These boxes are sometimes termed the *Voronoi polygons*. Using these we construct a new triangulation Π_h of the domain Ω, divide each of the Voronoi-polygons into triangles in such a way that the vertices of these triangles are the centre of the polygon and two neighbouring extremal points on the boundary of the polygon. Thus, two of the three vertices of these new triangles are points of intersection of midperpendiculars.

We now consider the mixed finite element method which makes use of the lowest-order RT-elements for the approximation of the fields and current densities on the triangulation Π_h, and of the piecewise constant approximations of the potentials on a box. In order to analyse this method, we introduce the following notation. For each mesh point (x_i, y_i) we let B_i be the box around it and T_i^k, $k = 1, ..., n_i$ the triangles of Π_h which are contained in B_i. Thus, we have that $\Pi_h = \cup_i \cup_{k=1}^{n_i} T_i^k$. We define

$$V_h = \{u_h \text{ is constant on each box}\}$$

This space is spanned by the basis functions ϕ_i, which are 1 on B_i and 0 on all other boxes. Remark that these functions are not continuous over the box edges. Furthermore, we let W_h be the space of lowest-order Raviart-Thomas elements on the triangles of Π_h, i.e. W_h is the space spanned by the vector basis functions r_j, which have a constant normal component equal to 1 along edge j of a triangle in Π_h and zero normal component along each of the other edges. Thus, the support of r_j extends over two triangles. On each of these triangles, r_j is of the form

$$r_j(x,y) = \frac{\pm l_j}{2area(T)} \begin{pmatrix} x - x_{(j)} \\ y - y_{(j)} \end{pmatrix}$$

in which $(x_{(j)}, y_{(j)})$ is the vertex of the triangle T opposite the edge j, and l_j the length of edge j.

We will now show that the box method is equivalent to this mixed finite element method when a suitable quadrature is chosen. To this end, set up the mixed FEM equation of the form (5) for the box B_i:

$$\int\int_{B_i} \nabla \cdot \mathbf{J}_h dO = \int\int_{B_i} R(u_h) dO \qquad (7)$$

Using Green's theorem, the fact that $u_h = u_i$ on B_i, and the property of the lowest-order RT-elements that their normal components along an arbitrary line are constant, we obtain:

$$\sum_{k=1}^{n_i} l_i^k J_i^{k,out} = area(B_i) R(u_i) \qquad (8)$$

where l_i^k is the length of the edge of T_i^k which coincides with the outer edge of the box B_i and $J_i^{k,out}$ is the normal component of $\mathbf{J}_h$ along that edge.

Next we use equation (6) for the r_j which correspond to edges which coincide with the outer edges of B_i. For each triangle $T_i^k \in B_i$ there is (except at the boundary) another triangle $T_{i'}^{k'} \in B_{i'}$ which, together, form the support of r_j. Then we have:

$$\int\int_\Omega a^{-1}\mathbf{J}_h \cdot r_j dO = \int\int_{T_i^k \cup T_{i'}^{k'}} a^{-1}\mathbf{J}_h \cdot r_j dO$$

$$\sim \frac{1}{dist(i,i')} \left\{ \int_{(x_i,y_i)}^{(x_{i'},y_{i'})} a^{-1} ds \right\} \mathbf{J}_h\left(\frac{x_i + x_{i'}}{2}, \frac{y_i + y_{i'}}{2}\right) \cdot \int\int_{T_i^k \cup T_{i'}^{k'}} r_j dO$$

The latter integral can be shown to be equal to

$$\frac{\pm l_j}{3} \begin{pmatrix} x_{i'} - x_i \\ y_{i'} - y_i \end{pmatrix}$$

Since T_i^k and $T_{i'}^{k'}$ are congruent, it then follows that

$$\int\int_\Omega a^{-1}\mathbf{J}_h \cdot r_j dO \sim \pm \left\{ \int_{(x_i,y_i)}^{(x_{i'},y_{i'})} a^{-1} ds \right\} \frac{l_i^k}{3} J_i^{k,out}$$

The treatment of the right hand side of (6) is slightly more complicated. Since the r_j have constant divergence on each triangle, we may replace the right hand side of the weak formulation:

$$\int\int_{T_i^k \cup T_{i'}^{k'}} u \nabla \cdot r_j dO \sim l_i^k \left(u(z_i^k) - u(z_{i'}^{k'})\right)$$

where z_i^k and $z_{i'}^{k'}$ are the centres of gravity of T_i^k and $T_{i'}^{k'}$, respectively. Using interpolation we obtain:

$$\int\int_{T_i^k \cup T_{i'}^{k'}} u \nabla \cdot r_j dO \sim \frac{l_i^k}{3} \left(u(x_i, y_i) - u(x_{i'}, y_{i'})\right)$$

Thus, the right hand side of (6) can be approximated by

$$\frac{l_i^k}{3}(u_{i'} - u_i)$$

Combining the obtained approximations, we finally get.

$$J_i^{k,out} = \frac{u_{i'} - u_i}{\int_{(x_i,y_i)}^{(x_{i'},y_{i'})} a^{-1} ds} \qquad (9)$$

Equations (8) and (9) yield the box scheme discretisation 1. Thus, we have shown that the latter can be obtained by approximating the integrals in the lowest order RT-method described above.

4. INTERPOLATIONS FOR THE FIELDS AND CURRENT DENSITIES OBTAINED USING THE BOX METHOD

The equivalence of the box method and a low-order mixed finite element method opens possibilities for obtaining expressions for the fields and current densities inside the elements (remember that the box method only provides components of these on the edges). We will describe two ways of doing this.

We can use equations (5) and (6) for the test functions which have not yet been used in the above. More specifically: for each box B_i, we can set up a system of n_i equations for the n_i remaining unknowns, namely the normal components of $\mathbf{J}_h$ along the inner edges of the box. In the resulting set of equations, the values of the components of $\mathbf{J}_h$ along the outer edges of B_i occur, for which we can substitute expression (9), as well as the values u_i which have already been determined. Remark that these calculations can all be done *locally*, i.e. this can be considered as a postprocessing exercise. Having obtained the values for the n_i remaining unknowns for the box B_i, we can give an expression for $\mathbf{J}_h$ inside the box: thus we have produced an $H(div;\Omega)$-function (with continuous normal components over the edges of the triangles T_i^k). This is important for adaptive runs, or for applications in which the fields and/or the current densities are coefficients in another equation (e.g. the temperature equation, or the hydrodynamic equations). In the lecture we will give an example of this.

The method just described for obtaining the fields and current densities inside the box does guarantee current conservation on the boxes, but not on the triangles of Π_h. The latter can be achieved by considering a mixed finite element method on each box B_i, in which we now introduce piecewise constant potentials on the triangles of Π_h. Thus, there are $2n_i$ remaining unknowns per box. In order to have a well-determined system of equations for these, we impose the extra restriction that the average of the newly introduced u_i^k (value of potential on T_i^k) is equal to the value u_i calculated in the centre of the box:

$$u_i = \frac{\sum_{k=1}^{n_i} area(T_i^k) u_i^k}{area(B_i)}$$

This second way of interpolating the results obtained by the box method does guarantee current conservation on the triangles of Π_h, and yields another approximation for the potentials.

References

[1] T.J.R. Hughes, A.N. Brooks *A multidimensional upwind scheme with no crosswind diffusion*, Finite element methods for convection dominated flows, T.J.R. Hughes (ed.), ASMC, New York, pp. 19-35 (1979)

[2] F. Brezzi, L.D. Marini, P. Pietra, *Two-dimensional exponential fitting and applications to semiconductor device equations*, Publ. no. 597, Consiglio Nazionale Delle Ricerche, Pavia, Italy (1987)

[3] S.J. Polak, W.H.A. Schilders, H.D. Couperus, *A finite element method with current conservation*, Proc. SISDEP-3 Conf., G. Baccarani and M. Rudan (eds.), Bologna, Italy, pp. 453-462 (1988)

[4] F. Brezzi, L.D. Marini, P. Pietra, *Mixed exponential fitting schemes for current continuity equations*, Proc. NASECODE-VI Conf., J.J.H. Miller (ed.), Dublin, Ireland, pp. 546-555 (1989)

[5] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers*, RAIRO, R 2, pp. 129-151 (1974)

[6] P.A. Raviart, J.M. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, Mathematical Aspects of the finite element method, Lecture Notes in Math. no. 606, pp. 292-315, Springer Verlag, Berlin (1977)

An approximate Newton-method for the Semiconductor Device Equations

Klaus Gärtner
Karl-Weierstraß-Institut für Mathematik
Mohrenstr. 39, D-O-1086 Berlin

Abstract - Solving the system of the stationary semiconductor device equations in three dimensions one is strongly interested to reduce the complexity of the problem. An approximate decoupling of the electrostatic potential and the carrier densities, motivated by singular perturbation theory, was introduced in [1]. This reduces the problem to a scalar one and a two times two system. In many (majority carrier) situations the changes in the minorities are mostly determined by the changes of the majorities - so that the problem is reduced to scalar problems. Typical convergence rates are 0.2,...,0.02 in cases where Gummel's method shows slow convergence (MOSFET with current but not very strong recombination).

I. APPROXIMATION FOR VAN ROOSBROECK'S EQUATIONS

The simplest version of van Roosbroeck's equations still fitting our purposes will be used - with a convenient scaling one arrives at the following system of Poisson's equation and continuity equations for electrons and holes:

$$
\begin{aligned}
&-\Delta u = f - n + p,\\
&-\nabla\cdot J_n = -R,\ J_n = \nabla n - n\nabla u,\\
&\nabla\cdot J_p = -R,\ J_p = -\nabla p - p\nabla u \ \text{ in } G,\\
&u = u^0,\ n = n^0,\ p = p^0 \ \text{ on } \Gamma_0,\\
&\nu\cdot u + \alpha(u-u^1) = \nu\cdot J_i = 0, i = n,p \ \text{ on } \Gamma_1.
\end{aligned}
\tag{1}
$$

Here: $G \subset R^d$, $1 \le d \le 3$, is a Lipschitzian domain with boundary $\Gamma = \Gamma_0 \cup \Gamma_1$, Γ_0, Γ_1 are disjoint, Γ_0 is closed and has a positive surface measure, ν is the outer unit normal at any point of Γ; the unknown functions u, n, p represent the electrostatic potential, the densities of electrons and holes, respectively; J_n and J_p denote the electron and hole current density; R the recombination, generation ratio; f is a given density of impurities; u^0, n^0, p^0 represent boundary values at ohmic (Dirichlet) contacts, u^1 and α are given functions modelling gate contacts. Einstein relations, relating mobilities and diffusion coefficients by a constant factor, are assumed, too. Let

$$
\begin{aligned}
&Au = f_1 + p - n,\ (f_1, h) := (f,h) + \int \alpha u^1 h \, d\Gamma_1,\\
&B(u)n = C(u)p = -R
\end{aligned}
\tag{2}
$$

be the discrete version of (1) with h as test function of the discretization. Let (U, N, P) be a given approximate solution of (2) and new variables x, y, z be introduced via $u = U + x$, $n = N(1+z-y)$, $p = P(1+z-x)$, so

$$
\begin{aligned}
&Au = f_1 + p - n,\\
&B(U)n + B'_u(N)(u-U) = -R(N,P) - R_n N(z-y) - R_p P(z-x),\\
&C(U)p + C'_u(P)(u-U) = -R(N,P) - R_n N(z-y) - R_p P(z-x)
\end{aligned}
$$

is the corresponding Newton linearized system. Motivated by [2] and the connected convergence problems in [1] an algorithm was suggested that should be understood as preconditioner decoupling the electrostatic potential and the carrier densities (with $\tilde{B}'_u = B'_u - R_p P$, $\tilde{B} = B + R_n$, $\tilde{C}'_u = C'_u + R_n N$, $\tilde{C} = C + R_p$, $\underline{P} = P1$, $\underline{N} = N1$):

$$
\begin{aligned}
(A+P+N)x - Ny - Pz &= f_1 + \underline{P} - \underline{N} - AU\\
&= F_1 + \underline{P} - \underline{N},\\
\tilde{B}'_u x + \tilde{B}Nx - \tilde{B}Ny + R_p Pz &= -B\underline{N} - R(N,P),\\
\tilde{C}'_u x - \tilde{C}Px + \tilde{C}Pz - R_n Ny &= -C\underline{P} + R(N,P).
\end{aligned}
$$

Choosing $\tilde{A}^{-1}$ as an approximate inverse of $A+N+P$ one has the following iteration scheme:

$$
\begin{aligned}
&(\tilde{B}'_u\tilde{A}^{-1} + \tilde{B}(N\tilde{A}^{-1} - E))Ny^{i+1} +\\
&((\tilde{B}'_u + \tilde{B}N)\tilde{A}^{-1} + R_p)Pz^{i+1} =\\
&-\tilde{B}\underline{N} - R(N,P) - (\tilde{B}'_u + \tilde{B}N)\tilde{f}_1,
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
&((\tilde{C}'_u - \tilde{C}P)\tilde{A}^{-1} - R_n)Ny^{i+1} +\\
&(\tilde{C}'_u\tilde{A}^{-1} + \tilde{C}(E - P\tilde{A}^{-1}))Pz^{i+1} =\\
&-\tilde{C}\underline{P} + R(N,P) - (\tilde{C}'_u - \tilde{C}P)\tilde{f}_1,
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
&(A+P+N)\ x^{i+1} =\\
&(F_1 + \underline{P} - \underline{N} + Ny^{i+1} + Pz^{i+1}),\ x^0 = 0.
\end{aligned}
\tag{5}
$$

$\tilde{A}$ is chosen as $\tilde{A} = a\ \mathrm{diag}[A] + N + P$, a damping parameter. $a = 0$ corresponds to singular perturbation theory, for $a \ge 1$ holds $\tilde{A}^{-1}_{ij} \le (A+N+P)^{-1}_{ij}$ and $a \to \infty$ is a modified linearized Gummel scheme. So in some neighbourhood of the thermodynamical equilibrium one has an 'interpolation' between two convergent (in the continuous case) methods. a is used to approximate a characteristic eigenvalue (α) of A in those (small) parts of the domain with $n + p \ll |f_1 + p - n|$ - or with other words in parts where the singular perturbed ansatz is violated and $n + p$

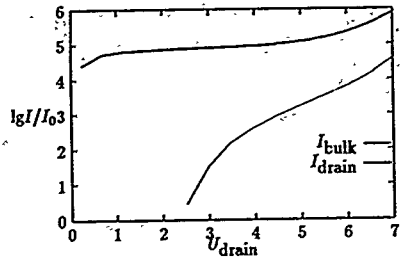


Figure1: Drain and bulk current with avalanche generation, $I_0 = 1.e-8A$, $U_{bulk} = U_{sour} = 0V, U_{gate} = 2V$, $\mu = \mu(n,p)$

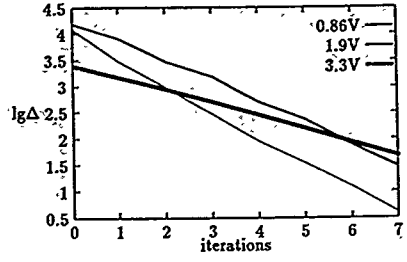


Figure2: Convergence ratio $\lg\Delta, \Delta = max(\frac{U_{i+1}-U_i}{U_0})$ of the ele. potential, $U_0 = 10^{-4}UT$, $U_{bulk} = U_{sour} = 0V$, $U_{drain} = 0.86, 1.9, 3.3V$, $U_{gate} = 2V$, $\mu = \mu(n,p)$, damping parameter $\alpha = 0.05$

should not be compared with 0 instead of α. (5) guarantees the same smoothness for x as for u and should not be simplified.

II. NUMERICAL RESULTS

The system (3) to (5) has been implemented in ToSCA (2d device simulation program of the Karl-Weierstraß-Institut für Mathematik) in the following manner: The remaining two times two system was solved by simple block forward backward substitution and one iteration starting with the majority carrier density and unmodified diagonal blocks $\tilde{B}'_n\tilde{A}^{-1} + \tilde{B}(N\tilde{A}^{-1} - E)$, $\tilde{C}'_n\tilde{A}^{-1} + \tilde{C}(E - P\tilde{A}^{-1})$. In cases without recombination and negative / positive definite B'_n / C'_n (valid for special discretizations) this blocks are regular. Not any density or field dependence of the mobilities was included in the Jacobian matrices and the tests have been performed for different MOSFETs. Results for an avalanche breakdown are shown here. Figure 1 shows the current voltage characteristics. Figures 2,3 give an impression of the influence of the damping parameter α at various bias conditions. Figure 4 illustrates the weak recombination coupling even at the highest injection levels and supports the diagonalization of the possibly two iteration processes. At lower current levels one iteration for solving (3) to (5) is optimal, in the avalanche region the minority carrier density must be calculated correctly and the singular perturbed ansatz is violated. Together with convergence acceleration three or four iterations should be sufficient to reach the error reduction defined by the outer approximate Newton process.

ACKNOWLEDGMENT - The author wants to thank Prof. Gajewski for many discussions and the support using ToSCA to test the algorithm.

REFERENCES

[1] H. GAJEWSKI AND K. GÄRTNER, *On the iterative solution of van Roosbroeck's equations.* ZAMM (To appear).

[2] C. RINGHOFER AND C. SCHMEISER, *An approximate Newton method for the solution of the basic semiconductor device equations,* SIAM J. Numer. Anal., 26 (1989), pp. 507–516.

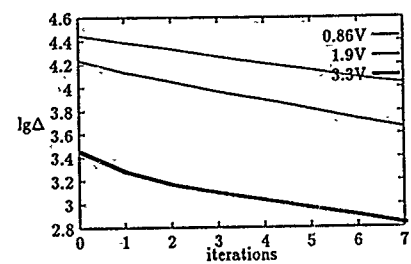


Figure3: Convergence ratio lgΔ of the ele. potential, $U_0 = 10^{-4}UT$, $U_{bulk} = U_{sour} = 0V$, $U_{drain} = 0.86, 1.9, 3.3V$, $U_{gate} = 2V$, $\mu = \mu(n,p)$, damping parameter $\alpha = 100$

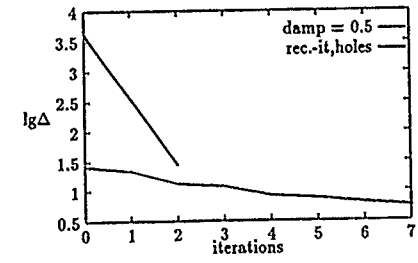


Figure4: Convergence ratios: inner iteration process for holes and eletrons (rec.-it,holes) and outer one for the ele. potential (damp=0.5)), $U_{bulk} = U_{sour} = 0V$, $U_{drain} = 0.86, 1.9, 3.3V$, $U_{gate} = 2V$, $\mu = \mu(n,p)$, damping parameter $\alpha = 0.5$

THREE-DIMENSIONAL TRANSIENT DEVICE SIMULATION WITH MINIMOS

OTTO HEINREICHSBERGER AND SIEGFRIED SELBERHERR
Institute for Microelectronics
Gußhausstraße 27-29
Technical University Vienna, A-1060 Vienna/Austria

Abstract

Our device simulator MINIMOS has been used for the numerical analysis of three-dimensional non-planar silicon MOSFET and GaAs MESFET structures. Here we present an extension of the program for the simulation of transient effects. This version of MINIMOS has further been enhanced by a new, highly accurate current integration method.
The computational complexity of three-dimensional transient simulations is tackled by preconditioned iterative methods. We present efficient algorithms and their implementation for the solution of the large linear systems of equations on vector and parallel computers.

1 Transient Simulation

Three-dimensional transient simulation of MOSFET structures is necessary to analyze both the influence of time dependent physical quantities such as the recombination rate, and three-dimensional non-planar geometries such as the field-oxide transition. The current continuity equations are discretized in space by the Scharfetter-Gummel method and in time by the fully implicit (backward Euler) method. Time-step control is based on the functional

$$I = \int_\Omega [(n^{i+1}-n^i)\ln\frac{n^{i+1}}{n^i} + (p^{i+1}-p^i)\ln\frac{p^{i+1}}{p^i} + \frac{\epsilon}{2}(\mathrm{grad}\,(\psi^{i+1}-\psi^i))]d\Omega$$

The time-step τ_{i+1} is chosen such that

$$\frac{\tau_{i+1}}{\tau_i} = \delta \cdot I$$

remains bounded. For the solution of the device equations the decoupled (Gummel) algorithm is used. The convergence of Gummel's algorithm in the linear regime is accelerated effectively by least squares extrapolation for the update of the electric potential ψ.

Transient simulation is an important method for the analysis of physical effects such as e.g. the kinetic of deep traps in the semi-insulating substrates of GaAs MESFETs [3][4], and the simulation of the charge-pumping experiment (interface trap kinetic). The deep trap model in GaAs for the donor trap rate-equation is given by

$$\frac{\partial\,(N_T - N_T^+)}{\partial t} = R_n - R_p$$

in which the effective recombination rates $R_{n,p}$ are

$$\begin{aligned} R_n &= C_n N_T^+ n - e_n\,(N_T - N_T^+) \\ R_p &= C_p\,(N_T - N_T^+)\,p - e_p N_T^+ \end{aligned}$$

In these equations N_T denotes the total and N_T^+ the density of the electrically active deep donors. $C_{n,p}$ are capture coefficients and $e_{n,p}$ are the emission rates [3]. An equivalent formulation holds for the acceptor traps. For the simulation of the charge-pumping experiment in silicon MOSFETs we use a model for the interface trap kinetic given in [2]. Assuming the acceptor type of the interface traps with density D_T, the time dependent charge-pumping current is obtained by integration with respect to the gate-oxide surface and the time (period length T_0). The falling pulse slope is assumed to start at $t = t'$:

$$I_{CP}(t) = \frac{1}{T_0}\sigma v_{th}\int_{t'}^{t'+T_0}\int_{\Gamma_G} N_T(\tau)p(\tau)d\Gamma d\tau$$

where $N_T(t)$ is the non-equilibrium part of the trapped charge density, obtained by integration of the rate equation

$$N_T(t) = -\int_{E_F(t')}^{E_F(t)} D_T(E)\exp\left(-\frac{\delta(t,E)}{\tau_n(E)}\right)dE - \sigma v_{th}\int_{t'}^{t} N_T(\tau)\,p(\tau)\,d\tau$$

The time interval $\delta(t, E)$ is determined for a given E by the condition $E_F(t-\delta) - E = 0$. All parameters are assumed as spatial variables along the channel surface. $\tau_n(E)$ denotes the trap lifetimes which depend on the energy in the well-known way [6].

A selfconsistent transient solution of these equations enables the simulation of the charge-pumping experiment. This facilitates a proper design of this experiment and the extraction of the spatial distribution and energy density of the traps created by a hot carrier injection.

2 Current Integration

After a solution at some timestep has been found a critical step is the terminal current integration. We have implemented a new method which is based on choosing weight functions w_n^i for each terminal T_i and evaluating a volume integral instead of a surface integral. E.g. for the electron current on the terminal T_i we compute

$$I_n^i = \int_\Omega [\mathrm{grad} w_n^i \cdot J_n - w_n^i R] d\Omega$$

In this formula J_n denotes the electron current density and R the recombination rate. The functions w_n^i are smooth functions on Ω. They have to suffice homogenous Dirichlet boundary conditions at all terminals $T_j \neq T_i$, non-homogenous constant Dirichlet boundary conditions at the terminal T_i and homogenous Neumann boundary conditions elsewhere. To obtain optimal weight function for e.g. the electron current on terminal T_i we minimize the functional

$$\Phi_n^i = \int_\Omega \left[\frac{n}{2}(\mathrm{grad} w_n^i)^2 - w_n^i R\right] d\Omega$$

This choice is motivated by the experimental observation that a high degree of accuracy for terminal currents is achievable, if the gradients of the weight functions are minimized in highly doped regions of the device [5]. The variation of the functional above leads to the elliptic partial differential equation

$$\mathrm{div}\,(n \mathrm{grad} w_n^i) = R$$

which is discretized by the Scharfetter-Gummel method and solved by the standard preconditioned conjugate gradient method. For the deviation currents the Laplace equation is solved on Ω.

3 Implementations

The backward Euler time discretization in general increases the diagonal dominance of the linearized discrete carrier continuity equations, thus making preconditioned iterative methods converge quickly. Apart from the classical conjugate gradient algorithm (CG), which is used to solve the discrete Poisson and weight function equation, we use the conjugate gradient squared method (CGS) for the carrier continuity equations.

The convergence rate on one hand and the efficiency of the implementation on parallel computers on the other is determined to a large extent by the applied preconditioner. Incomplete LU factorizations have proven to be a nearly optimal choice on vector computers. We have carried out various implementations on vector supercomputers such as the Cray-2 and the Fujitsu VP200 resulting in execution speeds of more than 100 megaflops for the critical triangular solves of the ILU preconditioner. This is achieved by the hyperplane-reordering technique using list-vectors. We have investigated also several massively parallelizable preconditioners, namely truncated Neumann series and multicolor incomplete factorization preconditioners. Experiments performed on a massively parallel architecture, the Connection Machine [1], indicate that an incomplete factorization of the reduced system matrix performs best.

References

[1] Heinreichsberger, O., "MINIMOS on the Connection Machine", Technical Report, Institute for Microelectronics, Tech. Univ. Vienna, Feb. 1991.

[2] Hofmann, F., Hänsch, W., "The Charge Pumping Method: Experiment and Complete Simulation", *J.Appl.Phys.* Vol. 66, 1989.

[3] Kazushige, H., Yanai H., Toshiaki I., "Numerical Simulation of GaAs MESFETs on the Semi-Insulating Substrate Compensated by Deep Traps", *IEEE-ED 35* No. 11, Nov. 1988.

[4] Lindorfer, Ph., "Numerische Simulation von GaAs MESFETs", PhD. Thesis, Techn. Univ. Vienna, 1991.

[5] Nanz, G., "Numerische Methoden in der zweidimensionalen Halbleitersimulation", PhD. Thesis, Techn. Univ. Vienna, 1989.

[6] Sze, S.M., "Physics of Semiconductor Devices", Wiley, ISBN 0-471-09837-X, 1986.

3D NUMERICAL SIMULATION OF STEADY-STATE AND TRANSIENT PROCESSES IN SILICON SEMICONDUCTOR DEVICES

A.I.ADAMSONS, B.S.POLSKY AND A.I.SHUR

Research Institute of Mathematics and Computer Science, University of Latvia,
29 Rainis Boulevard, Riga, U.S.S.R.

Abstract - Numerical methods for 3D steady-state and transient simulation of semiconductor devices are considered. These methods are realized in the devices simulator "ALPHA-3", which is the extension of our previous 2D simulator "ALPHA".

I. INTRODUCTION

The development of future semiconductor devices with submicron feature size requires the use of mathematical modeling for the analysis of charge transport in these devices and the estimation of their electrical characteristics. The behaviour of a semiconductor device is described by a nonlinear system of partial differential equations (drift-diffusion model). In general case these equations cannot be solved analytically without very hard simplified assumptions such as one-dimensional approach, constant mobilities, charge neutrality, trivial doping profiles and so on. Since for modern devices these assumptions are not valid, therefore numerical technique is the only tool for obtaining the solution.

Effective numerical methods and related software for 1D and 2D simulation are developed in the present time. On the other hand for small devices 3D effects may be quite essential as first results of 3D analysis show [1-6]. In the present paper our experience in the field of 3D simulation is given. The numerical methods for steady-state and transient analysus realised in the silicon device simulator "ALPHA-3" are described.

II. MATHEMATICAL MODEL

The physical behaviour of a semiconductor device is described by the following drift-diffusion model:

$$div\,\vec{J}_n = R(\rho,n) + \frac{\partial n}{\partial t}, \qquad (1)$$

$$div\,\vec{J}_p = -R(\rho,n) - \frac{\partial \rho}{\partial t}, \qquad (2)$$

$$\Delta\varphi = n - \rho - N, \quad N = N_d - N_a, \qquad (3)$$

$$\vec{J}_n = \mu_n(\nabla n - n\nabla\varphi - n\nabla \ell n n_{ie}), \qquad (4)$$

$$\vec{J}_p = -\mu_p(\nabla p + p\nabla\varphi - p\nabla \ell n n_{ie}), \qquad (5)$$

$$\vec{J}_d = -\frac{\partial}{\partial t}\nabla\varphi, \qquad (6)$$

$$\vec{J} = \vec{J}_n + \vec{J}_p + \vec{J}_d \qquad (7)$$

The system is written in normalized form and symbols have their standard meaning. The expressions for mobilities μ_n and μ_p according to Caughey-Thomas model are used. The recombination rate R includes three mechanisms: Shockley-Read-Hall process, Auger process and the cumulative multiplication of free carriers under strong electric field. The effective intrinsic concentration n_{ie} depending on the bandgap narrowing is calculated according to the formula of Slotboom or may be obtained from the solution of the special equalibrium problem [7].

III. NUMERICAL METHODS AND PROGRAM IMPLEMENTATION

The numerical solution of drift-diffusion equations is carried out by a finite difference method. For electron and hole current components Scharfetter-Gummel approximations [8] and the standard seven-point approximation of Laplace operator are used. Steady-state problems are solved by decoupled Gummel's method [9], in which linearized Poisson equation and continuity equations are iterated successively (external iterative process) until the self-consistent solution is obtained. Each Gummel's iteration requires the solution of three linear systems of elliptical difference equations. For this purpose iterative methods [10, 11] are used (internal iterative process).

Absolutely stable half-implicit scheme [12] for the analysis of transient processes is used. In this scheme the determination of the potential of electric field is carried out in two stages. Firstly the total current continuity equation, which is the consequence of (1)-(7), is solved and the predictor for the potential is obtained. The final value of the potential is determined from the stabilized Poisson equation. The automatical time step selection with local error control is realized.

Above mentioned methods are implemented in the device simulator "ALPHA-3", which is the extension of our 2D simulator "ALPHA" [7]. CPU time for one Gummel iteration on the grid with 28×28×18 knots is approximately 3 min on the soviet mainframe ES-1060 (1mips). Depending on the injection level from 5 to 50 Gumel iterations are reguired for the calculation of one point of I-V curve. CPU time for one time step in the transient analysis is also 3 min and 200-300 steps are sufficient for the simulation of transient process with the tolerance 10^{-2}.

REFERENCES

1. Yoshii A., Kitazawa H., Tomizawa N. et al, "A three-dimensional analysis of semiconductor devices", IEEE Trans., vol.ED-29, Nr.2, P.184-189, 1982.
2. Adamsone A.I. and Polsky B.S., "Three-dimensional numerical simulation of bipolar transistor devices", Radioelectronica, vol.29, Nr.9, P.46-49, 1986 (in Russian).
3. Adamsone A.I. and Polsky B.S., "Numerical modelling of submicrometer bipolar transistor", Radioelectronica, vol.32, Nr.6, P.74-75, 1989 (in Russian).
4. Odanaka S., Hiroki A., Umimoto H. et al, "SMART-II: A three-dimensional CAD model for submircometer MOSFET's", Proc. NASECODE VI Conf., Dublin: Boole Press, P.303-310, 1989.
5. Toyabe T., Masuda H., Aoki Y. et al, "Three-dimensional device simulator CADDETH with highly convergent matrix solution algorithms", IEEE Trans., vol.ED-32 Nr.10, P.2038-2044, 1985.
6. Buturla E., Johnson J., Furkay S. et al, "New 3D device simulation formulation", Proc. NASECODE VI Conf., Dublin: Boole Press, P.291-295, 1989.
7. Polsky B.S., "Numerical simulation of semiconductor devices", Riga, Zinatne, 168p. 1986 (in Russian).
8. Scharfetter D.L. and Gummel H.K., "Large-signal analysis of a silicon Read diode oscilator", IEEE Trans., vol.ED-16, Nr.1, P.64-77, 1969.
9. Gummel H.K., "A self-consistent iterative scheme for one-dimensional steady-state transistor calculations" IEEE Trans., vol.ED-11, Nr.10, P.455-465, 1964.
10. Meijerink J.A., Van der Vorst H.A., "An iterative solution method for linear system of which the coefficient matrix is a symmetric M-matrix", Math. of Comp., vol. 31, Nr.137, P.148-162, 1977.
11. Buleev N.I. "Incomplete factorization method for solution of two- and three-dimensional diffusion equations", ZVM i MF, vol.10, Nr.4, P.1042-1044, 1970 (in Russian).
12. Polsky B.S. and Rimshans J.S. "Half-implicit difference scheme for numerical simulation of transient processes in semiconductor devices", Solid-State Electron, vol.29, Nr.3, P.321-328, 1986.

COMPUTER AIDED ANALYSIS OF METASTABLE DECAY REACTIONS OF IONIZED VAN DER WAALS CLUSTERS

P. Scheier and T.D. Märk

Institut für Ionenphysik, Universität Innsbruck, A-6020 Innsbruck, Austria

Abstract: The computer aided mathematical treatment of two problems connected with the metastable decay of ionized Van der Waals (vdW) clusters will be presented here. One is the mathematical treatment and analysis of sequential decay series in nitrogen clusters in order to allow the meaningful determination of decay rates. Secondly, we developed a computer program to facilitate the correct interpretation of metastable cluster ion spectra allowing to exclude after identification all possible coincidences.

Introduction: Small atomic and molecular aggregates (clusters) are of interest both from the point of view of basic and applied science. Clusters are seen as a link between the gas phase and the condensed state. Van der Waals and especially rare gas clusters are rather simple systems and therefore also accessible to theoretical treatments. Careful observation and (mathematical) analysis of metastable decay reactions of ionized vdW clusters allows to draw conclusions about stability, structure and internal energy transfer in these systems.

Experimental Method: Neutral clusters are produced by expanding pure or seeded gas through a small nozzle into a vacuum chamber. The ensuing supersonic beam passes a skimmer and is crossed 10 cm downstream at right angles by an electron beam of variable energy. Ions thereby produced are extracted at right angles to the plane of the neutral cluster beam and electron beam, accelerated to a typical energy of 3 keV and mass selected in a high resolution double focussing mass spectrometer system.

Important for the present studies of metastable dissociations is the existence of two field free regions. The first one between the end of the acceleration and the entrance slit of the magnetic field and the second field free region between the two sector fields. A possible metastable decay of an ion (mass m_1,charge q_1) into (m_2,q_2) can be detected either in the first field free region by tuning the magnetic sector field to a nominal mass $m^* = (m_2^2/q_2^2)/(m_1/q_1)$ and the electric sector field to a voltage $E^* = E(m_2/q_2)/(m_1/q_1)$ (with E the voltage of the electric analyzer for direct mass spectra) or in the second field free region by tuning the magnetic field to m_1/q_1 and the electric sector field to $E^* = E(m_2/q_2)/(m_1/q_1)$. In the present study however, we used an additional technique to study the occurrence of successive metastable decay series. In this alternative mode of operation it is possible to detect a decay of (m_1,q_1) into (m_2,q_2) in the first field free region followed by a sequential metastable decay of this (m_2,q_2) into (m_3,q_3) in the second field free region by tuning the magnetic sector field to a nominal mass $m^* = (m_2^2/q_2^2)/(m_1/q_1)$ and the electric sector field to an electric field $E^* = E(m_3/q_3)/(m_1/q_1)$.

Results: Magnera et al /1/ recently reported that N_2-cluster ions do not only lose one monomer, but depending on the cluster size they observed well structured decay patterns. For example $(N_2)_{45}^{+*}$ cluster ions are loosing with high probability 2,7,12,17 and 22 monomers. They could not determine whether these monomers are lost in one step or sequentially. With the above mentioned technique we were able to demonstrate, that all of these decays are sequential /2/.

The next step was to develope a program which allows to evaluate from the experimental data apparent decay rates. The following set of coupled differental rate equations are describing these decays:

$$dN_0(t)/dt = -k_0N_0(t)$$
$$dN_1(t)/dt = k_0N_0(t) - k_1N_1(t)$$
$$\vdots$$
$$dN_i(t)/dt = k_{i-1}N_{i-1}(t) - k_iN_i(t)$$

The solution of these equations is very simple leading to exponential dependencies, but the actual calculation of the decay rates is only possible with a numerical method. For example for the $(N_2)_5^{+*}$ the following decay channels can be measured:

5 mothersignal
5 -1-> 4, 5 -2-> 4,
5 -1-> 3, 5 -2-> 3, 5 -1-> 4 -2-> 3
5 -1-> 2, 5 -2-> 2, 5 -1-> 4 -2-> 2, 5 -1-> 3 -2-> 2
5 -1-> 1, 5 -2-> 2, 5 -1-> 4 -2-> 1, 5 -1-> 3 -2-> 1, 5 -1-> 2 -2-> 1

(where 5 -1-> 4 stands for a decay of $(N_2)_5^+$ into $(N_2)_4^+$ in the first field free region, and 5 -1-> 4 -2-> 2 stands for a sequential decay where $(N_2)_5^+$ decays to $(N_2)_4^+$ in the first field free region followed by a decay of this $(N_2)_4^+$ into $(N_2)_2^+$ in the second field free region).

In general for a n-mer $^1/_2n(n+1)$ experimental data are therefore available for the determination of n variables. Every measured decay channel is leading to one exponential equation, i.e., in the 5-mer case leading to 15 equations for only 5 variables. Nevertheless, using a combination of a Gaussian with a Monte Carlo method it was possible to obtain decay rates, which describe the experimental data very well, i.e., using the derived decay rates it was possible to calculate the ion signals for every decay channel in good agreement with the experimental result. Table 1 shows the results in case of $(N_2)_3^+$

Another problem in clusterphysics is the occurrence of coincidences. In direct mass spectra of analytical chemistry coincidences are a well known problem. In particular when a sample consists of many different hydrocarbons, the mass

spectrometric fragmentation patterns of the organical molecules are very much overlapping /3/. The situation becomes even worse if one is using clusters. Table 2 shows as an example some of the coincidences of propane and mixed CO_2-H_2O clusters at a nominal mass of 100 Daltons.

For the identification of metastable peaks (which is even more difficult than direct spectra) in magnetic sector-field mass spectrometer systems up to now a book by Beynon and coworkers /4/ could be used containing a list of all possible decays up to a mass of 500 Daltons. Besides the fact that it is not very convenient to search for a specific decay in this book, the maximum mass of 500 Daltons is much too small for the use with clusters. We have developed a new computer program (CLUSTER) which allows to calculate all the possible decays for a certain cluster series (also mixed clusters and multiply charged ones) and to list them in a file or on a printer. This program also takes into consideration all the possible fragmentations of molecular monomers For instance, for mixed CO_2-H_2O clusters there are 9562476 possible metastable decays of cluster ions up to a mass of 1000 Daltons Table 3 illustrates for several cluster ions possible coincidences as given on the printer readout of this new program (CLUSTER).

Acknowledgement: Work supported partially by the Österreichischer Fonds zur Förderung der Wissenschaftlichen Forschung.

References:

/1/ T.F. Magnera, D.E. David and J. Michl; Chem. Phys. Lett. 123 (1986) 327

/2/ P. Schéler and T.D. Märk; Phys. Rev. Lett. 59 (1987) 1813

/3/ C. Brunnee and H. Voshage; Verlag Karl Thiemig KG München (1964) 69

/4/ J.H. Beynon, R.A. Saunders and A.E. Williams; Elsevier (1965)

Table 1: Comparison of metastable fractions of the measured decays of $(N_2)_3^+$ clusterions and the calculated data using the evaluated decay rates.

Decay channel	experimental metastable fraction	calculated metastable fraction	difference
3 -1-> 3 -2-> 2	1.41252 %	1.41312 %	0.00060 %
3 -1-> 3 -2-> 1	0.00674 %	0.00376 %	0.00298 %
3 -1-> 2 -2-> 2	8.47512 %	8.47537 %	0.00025 %
3 -1-> 2 -2-> 1	0.00433 %	0.00161 %	0.00272 %
3 -1-> 1 -2-> 1	0.08989 %	0.05614 %	0.03375 %

Table 2: Coincidences of propane and mixed CO_2 H_2O-clusters in the direct spectrum at a nominal mass to charge ratio of 100 Daltons.

from $(C_3H_8)_n$	from $(CO_2)_i(H_2O)_j$
$(C_3H_8)_2C$	$(CO_2)_2C$
$C_3H_8(CH_2)_4$	$(H_2O)_4CO$
$C_3H_8CH(CH_2)_2CH_3$	$CO_2(CO)_2$
$C_3H_8(CH_2(CH_3)_2$	$H_2OO_3(OH)_2$
$C_3H_8CCH_2(CH_3)_2$	$H_2OC_4(OH)_2$
$(CH_2)_5(CH_3)_2$	$(H_2O)_2C_3CO$
⋮	⋮
7207 coincidences	5227 coincidences

Table 3 Coincidences in metastable windows Δm = 5 Daltons, ΔE = 5 Volt, maximum mass 1000 Daltons.

nominal mass m^* = 200 Daltons, nominal sector field E^* = 400 Volt:

$Ar_n(O_2)_m$:	example	m^*	E^*
	$Ar_7O_2O^+$ -1-> Ar_6O^+	199.805	398.829
	$Ar_5(O_2)_4^+$ -1-> $Ar_4(O_2)_3^+$	199.805	398.829
	in total 4 coincidences		
Ar_nq^+ (q<4):	example	m^*	E^*
	Ar_{24}^{3+} -1-> Ar_{19}^{3+}	200.556	404.542
	Ar_{25}^{3+} -1-> Ar_{13}^{2+}	202.800	398.580
	in total 2 coincidences		
$(C_3H_8)_n$:	example	m^*	E^*
	$(C_3H_8)_4C_2H_4^+$ -2-> $(C_3H_8)_3C_2H_4^+$	204.000	400.784
	$(C_3H_8)_7C_2H_4^+$ -1-> $(C_3H_8)_5C_3H_4^+$	201.190	395.417
	in total 43 coincidences		

nominal mass m^* = 500 Daltons, nominal sector field E^* = 400 Volt.

$Ar_n(O_2)_m$:	example	m^*	E^*
	$Ar_{20}O^+$ -1-> Ar_{16}^+	501.961	400.784
	$Ar_7(O_2)_2^+$ -2-> $Ar_7(O_2)_3O^+$	504.000	397.444
	in total 102 coincidences		
Ar_nq^+ (q<4):	example	m^*	E^*
	Ar_{41}^{2+} -1-> Ar_{16}^+	499.512	398.829
	Ar_{61}^{3+} -1-> Ar_{32}^{2+}	503.607	402.098
	in total 8 coincidences		
$(C_3H_8)_n$:	example	m^*	E^*
	$(C_3H_8)_{11}C^+$ -2-> $(C_3H_8)_8C_3^+$	496.000	399.734
	$(C_3H_8)_{18}C_2H^+$ -1-> $(C_3H_8)_{14}C_2H^+$	502.912	400.919
	in total 96 coincidences		

nominal mass m^* = 800 Daltons, nominal sector field E^* = 400 Volt.

$Ar_n(O_2)_m$:	example	m^*	E^*
	$Ar_{18}(O_2)_2O^+$ -2-> $Ar_{15}O_2^+$	800.000	403.690
	in total 40 coincidences		
Ar_nq^+ (q<4):	no coincidences in this window found		
$(C_3H_8)_n$:	example	m^*	E^*
	$(C_3H_8)_{18}C^+$ -2-> $(C_3H_8)_{14}C^+$	804.000	399.139
	in total 5 coincidences		

RANDOM NETWORK MODELS OF AMORPHOUS TETRAHEDRALLY BONDED SEMICONDUCTORS

F. WOOTEN*

Department of Applied Science, University of California, Davis, CA 95616, U.S.A.

and

D. WEAIRE†

Physics Department, Trinity College, Dublin 2, Ireland

Abstract. We describe a computer algorithm that generates models of covalently bonded amorphous semiconductors that are characterized by tetrahedrally coordinated random networks, the prototypical example being amorphous silicon. The starting point is a supercell in the diamond cubic structure. Periodic boundary conditions are imposed so as to eliminate surface effects. The algorithm uses an elementary topological rearrangement that is randomly and progressively introduced into the diamond cubic structure until a random network has been generated. This is followed by an annealing process that allows topological relaxation of the structure, leading to a structural model that is in good agreement with experiment as determined by the density and radial distribution function. We have recently modified the algorithm so as to generate networks having only, or mostly, even rings, as would be expected for a-GaAs. In anticipation of the likelihood that a few odd rings may remain after the annealing process, we have begun a program to quantify and visualize the topological defects.

I. INTRODUCTION

Covalently bonded amorphous semiconductors such as amorphous silicon (a-Si) are characterized as random network structures. Ideally such a structure is a well defined network of nearest-neighbor bonds, which may be analyzed in topological terms, for some purposes. An algorithm has been developed based on simulated annealing and bond-switching (illustrated in Fig. 1) that successfully models the homogeneous region that is believed to be representative of the bulk material in a-Si [1,2,3]. We are not attempting to directly model the physical process of equilibration, hence the method has much in common with the method of simulated annealing described by Kirkpatrick *et al* [4]. Such a model, and the process for its generation, has been given the name *sillium* [3,6]. The topology of sillium is characterized by a distribution of ring sizes, including a large fraction of odd rings.

There is considerable interest in extending the computer-modeling of amorphous structures to such covalently bonded materials as GaAs. For such binary compounds as GaAs there can be only a few "wrong bonds" (Ga-Ga or As-As), and in the ideal case there should be *no* wrong bonds and thus only even rings.

There are algorithms that ensure the generation of an even-ring random network starting from a network with only even rings such as the diamond-cubic structure, but the bond switching mechanism is a bit more complicated than that of Fig. 1 and, what is more important, it introduces much more strain into the network. One of these methods has been described by Rivier *et al* [5]. However, the

*Supported in part by the U.S. Department of Energy and Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

†Supported by EOLAS (Irish Science and Technology Agency).

Rivier model and *all* similar models that we have generated by computer over a period of some five years suffer from very large geometric distortions and are quite unphysical.

It seems that the best prospect for the computer generation of an even-ring model may be to return to the simplicity of the original bond-switch of Fig. 1 but modified by a weighting factor to bias against the creation of odd rings. The goal is the generation of an idealized model of the homogeneous bulk region of materials such as GaAs. On the other hand, even a model with just a few odd rings would be of interest, as the occasional inclusion of odd rings is not unexpected as a topological defect in real structures.

II. SILLIUM

The generation of sillium has been thoroughly described in earlier publications [2,3,6]. Here we present only a brief description of the essential ingredients.

The starting point is a supercell in the diamond cubic structure, subject to periodic boundary conditions, that contains enough atoms to provide a model of reasonable size. We have used supercells containing from 216 to 4096 atoms.

The starting crystal structure is then disordered by a sequence of randomly selected bond switches, as illustrated in Fig. 1. The bond switch illustrated there gives the simplest topological rearrangement that preserves tetrahedral bonding and introduces the minimum possible strain into the network. It also introduces 5- and 7-fold rings into the otherwise perfect diamond cubic structure. This process of introducing topological disorder is continued until there is no remaining memory of the original crystal structure.

The most stringent test of any remaining long range order is the structure factor $|S(q)|^2$ associated with those reciprocal lattice vectors labeled (111) for the diamond cubic structure [3,7]. The structure factor is initially of zero intensity for most values of the reciprocal lattice vectors for the chosen supercell. It is nonzero only for some q-values appropriate to the smaller unit cell of the diamond structure. The introduction of disorder redistributes the structure factor among all of the q-values of the supercell. We require that random bond switching continue until the structure factor is of roughly equal intensity for all q-values.

Every time two bonds are switched, the structure is relaxed by the method of Steinhardt *et al* [8] to the geometrical configuration that minimizes its energy as given by the Keating potential [9].

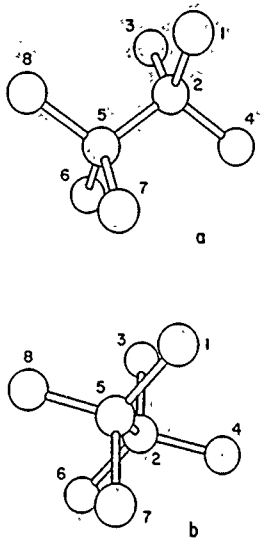


Figure 1: Local rearrangement of bonds used to generate random networks from the diamond cubic structure. (a) Configuration of atoms and bonds in the diamond cubic structure; and (b), relaxed configuration of atoms and bonds after switching bonds.

After the network has been randomized, further bond switches are randomly selected on a trial basis. The switches are accepted or rejected according to the usual Monte Carlo prescription: They are accepted if $\Delta E \leq 0$; they are accepted with probability $p = exp(-\Delta E/kT)$ if $\Delta E \geq 0$. (The Monte Carlo prescription is actually followed from the outset in most cases. That means the original randomization is carried out at a finite temperature [3] rather than the infinite temperature implied by the randomization process described above. This takes a bit more computer time for the randomiza-

tion process but leads to less highly strained randomized starting structures and reduces the total time for generating a satisfactory model.)

It frequently happens during the annealing process that a structure will be trapped in a metastable state from which no allowed bond switches will lead directly to a lower energy. The Boltzmann factor is essential in order to provide an escape mechanism from these local minima. It does this by occasionally allowing a transition to a somewhat more strained state from which another path to lower energy states can be found.

The structure is annealed by generating a sequence of networks while lowering the temperature in small increments. Typically, we require that at each temperature there be as many actual bond switches as there are bonds, in order to achieve equilibrium. The result is a model that, for silicon, *always* agrees very well with experiment as determined by the radial distribution function.

III. THE TRAVELING SALESMAN

It is illuminating that the algorithm we have developed for modeling a-Si can, with trivial modifications, be used to find a *reasonable* solution to the traveling salesman problem [10].

The simplest version of the traveling salesman problem is this: A traveling salesman must visit N cities and return to his home base by the shortest route. His possible routes of travel and distances between cities are given by a road map. The algorithm then treats the cities as atoms and the roads between cities as bonds between atoms. The energy in the Boltzmann factor becomes a path length and the temperature becomes a pseudo-temperature. (There is, of course, a question of how to interpret the temperature even when modeling a-Si, but the connection with a "real" temperature is at least qualitatively and conceptually direct for that case.)

Unlike the three-dimensional network needed to model a-Si, the closed path connecting N cities is two-dimensional. This makes it a simple matter to explicitly demonstrate a metastable state.

It happens that in the course of switching possible paths between cities a route is occasionally found such that, given the rules of the algorithm, there is no single switch in paths between cities that will further reduce the distance traveled in a complete circuit. Since there are $N(N-1)!/2$ possible routes, if N is large, the probability that the route is the minimum path is essentially zero. One has almost certainly found a metastable state! It is easy to depict such a state graphically [10].

IV. EVEN-RING MODELS

GaAs is the prototypical binary covalent semiconductor. In the idealized amorphous form it is assumed to be characterized by a random network having only even-fold rings; but, it is otherwise taken to be equivalent to sillium, that is, it has no dangling bonds, employs periodic boundary conditions, and is generated by the same simple bond switching mechanism.

Clearly something must be added to the prescription for generating models if only even-fold rings are wanted. We have attempted to accomplish this by including an additional term in the energy of the structure without modifying the Keating potential itself. The additional term is an effective energy that is proportional to the number of irreducible odd-rings. In practice, we have thus far counted only 5-fold rings. We express the energy as

$$E = E_{Keating} + \eta R_5$$

where R_5 is the number of 5-fold rings and η is a parameter to be optimized by (computer) experiment.

The only place that R_5 actually enters the modeling process is in the Boltzmann factor. It biases against the introduction of 5-fold rings. Put more positively, it favors the elimination of 5-fold rings that are necessarily introduced during the original randomization process.

We have chosen not to count 7-fold and higher rings for two reasons: It is very time consuming to count large rings; and, many higher odd-fold rings are reducible to 5-fold rings and thus disappear when the associated 5-fold ring is eliminated.

To understand how 7-fold reducible rings may disappear when a 5-fold ring is eliminated, consider what is meant by a reducible ring. If there are two atoms on the ring that are connected by a shortcut that does not lie on the ring, the ring is reducible. A typical example of a reducible ring is a 7-fold ring made from a 4-fold ring and a 5-fold ring that share a common bond. The outer 7-fold ring, consisting

of all the atoms from both the 4-fold and 5-fold ring and all the bonds from those rings except the one shared in common, is reduced by the common bond shared by the 4-fold and 5-fold ring. For that case, eliminating the 5-fold ring by, say, converting it to a 6-fold ring, also eliminates the reducible 7-fold ring by converting it to an 8-fold ring.

We have been successful in eliminating at least 85% of the 5-fold rings usually present in a random network structure like sillium. We believe it is possible to eliminate nearly all odd rings with further refinements in the algorithm.

V. RIVIER LINES

In anticipation of the likelihood that all odd rings may not be removed during the annealing process, and thus that it would be desirable to characterize the topology of the structure associated with the odd rings, we have simultaneously begun a program to quantify and visualize the topological defects. This program is being carried out in collaboration with J. Koch and will be described in detail in work to be published.

The essence of the program is that all irreducible rings in the structure are identified and triangulated. Each triangle is identified, numbered and stored along with the parity of the parent ring. A cubic grid of points is defined with a mesh scale finer than the structure of the random network. A search is then made for a path that lies on the grid and which passes through *only* odd rings and either makes a complete circuit, returning to the starting point, or extends to infinity. The latter is possible here because of the use of periodic boundary conditions. We refer to these "odd lines" as Rivier lines after their promoter N. Rivier [11].

We have found the two Rivier lines that thread their way through the four irreducible 5-fold rings and the four irreducible 7-fold rings that are created when a single pair of bond switches is introduced into the otherwise perfect diamond cubic structure. These lines are easily found by inspection and thus make a good first test case of the correctness of the computer algorithm. One of the lines circles the bond between atoms 1 and 5 in Fig. 1b. The other circles the bond between atoms 2 and 6 in Fig. 1b. Each Rivier line passes through two irreducible 5-fold rings and two irreducible 7-fold rings. We are now testing the program with more disordered structures, including the kinds of amorphous models whose generation has been discussed here.

VI. CONCLUSION

Historically, the study of random structures (e.g., liquids) has mainly concentrated on various distribution functions that express geometrical properties. In common with some recent problems in magnetism (spin glasses, etc.) the study of amorphous semiconductors has brought topological properties to the fore. While the search for direct probes of topologically defined quantities has proved frustrating in practice (despite several good ideas in principle), the topological approach to model-building and structure analysis remains a source of fascination to the theorist and stimulation to the experimenter. The work described here highlights the topological aspect, while respecting the constraints upon realistic geometrical arrangements.

REFERENCES

1. F. Wooten and D. Weaire, J. Non-Crystalline Solids 64 (1984) 325.
2. F. Wooten, K. Winer and D. Weaire, Phys. Rev. Letters 54 (1985) 1392.
3. F. Wooten and D. Weaire, Solid State Physics 40 (1987) 1.
4. S. Kirkpatrick, C. D. Gelatt, Jr. and M. P. Vecchi, Science 220 (1983) 671.
5. N. Rivier, D. Weaire and R. de Römer, J. Non-Crystalline Solids 105 (1988) 287.
6. J. Wejchert, D. Weaire and F. Wooten, J. Non-Crystalline Solids 122 (1990) 241.
7. F. Wooten and D. Weaire, J. Phys. C: Solid State Physics 19 (1986) L411.
8. P. Steinhardt, R. Alben and D. Weaire, J. Non-Crystalline Solids 15 (1974) 199.
9. P. N. Keating, Phys. Rev. 145 (1966) 637.
10. F. Wooten and D. Weaire, Key Engineering Materials 13-15 (1987) 109.
11. N. Rivier, Advances in Physics 36 (1987) 95.

A TOOL FOR PROCESS AND DEVICE MODELING INTEGRATED IN A DESKTOP ENVIRONMENT.

Yves MOREAU - Djamila SELIMI
Centre d'Electronique de MONTPELLIER
(Unité associée C.N.R.S. 391) - Case Courrier 084
Université MONTPELLIER Sciences (II)
- 34095 - MONTPELLIER CEDEX 5 (FRANCE)

☎ (33) 67 14 37 70
E-Mail: YMOREAU@FRMOP11.BITNET
Télécopie: 67 54 30 79

ABSTRACT

The development of semi-conductor devices involves several iterations of trial and error in the fabrication until a specified goal in terms of design conditions is reached. The application of device models can now decrease the number of trial and error steps. The reflection accompanying the development will be more productive if the modeling tools themselves do not require too much attention. We have developped a software package whose aim is to be used as easily as a usual word processor.

INTRODUCTION

New technologies applied to semi-conductor structures imply more and more complex behavior related to bi-dimensional effects and irregular profiles in miniaturized layers. Traditional analysis is often based on analytical expressions derived from solutions of simplified models (often in one dimension), and so, becomes less and less adapted. Numerical simulation appears as a precious tool for analysis and development of to-day's electronic components[1],[2],[3], its use definitely is becoming widespread, however, it appears that the closed form formulation is still widely used: two reasons can be invoked to explain this:

-(i) the physical interpretation based on the influence of each observable parameter is relatively easy to extract from a formula,

-(ii) the use of a software tool needs a time investment in learning which cannot be always afforded by the product designers.

Keeping in mind these reasons (*difficult man-machine relationship*), we have developed a software tool that can be run on a personal computer, its access being comparable to present word processors. It is implemented on an Apple MacIntosh™ computer and takes advantage of its editing facilities and user-friendly approach: *menus, windows, mouse "click and point"* operations...The user-friendly approach has to be developped on both sides of the numerical modeling : inputs (edition of the device) and outputs.

BASIC IDEAS

Rather than a coupling with a preprocessor and post processor (as in traditional procedural device simulators), it is a complete software with several mixed parts : edition of the device, 2-D simulation of diffusion processes and 2-D simulation of electrical behavior (electrostatic potential and current). Graphic control and questioning the computer is made available during all the phases of simulations. The basic mechanism is an "event loop" (as in usual productivity software) controlling the instantiation and the treatment of data structures. *Object-oriented programming* [1]ideas have guided the development of this software. Events are of different kinds : the traditional ones such as general messages to windows (mouse clicks to scroll, to move, to close...), drawing actions, menu-commands, user's answers to dialogs, as well as specific ones such as the end of computation of a local doping, or the convergence of the iterative computation of potentials and charges.

DEVICE EDITING

The description is made through the use of the mouse in "painting" the different zones of the studied device. This operation is consistent with the traditional work done with drawing softwares such as MacDraw™, Mac Paint™...the "colors" of the *palette* are here substrate N or P, oxide, contact diodes, gates ... or dopant implantations to allow precise description, scaling is provided and coordinates in nm are displayed while the user is drawing or modifying his zones. The zones are defined for the software, as polygonal objects with information on position, geometry, type, flags...and with specific data. All these objects are chained to form a list which is the studied device. The device is coupled with the window in which it is drawn and with a file where it can be saved for later use.

Figure 1 shows the computer screen with the palette, on the left, the drawing window, and three result windows (*List, plot net* windows) after computation of a local distribution of dopants (see next paragraph)

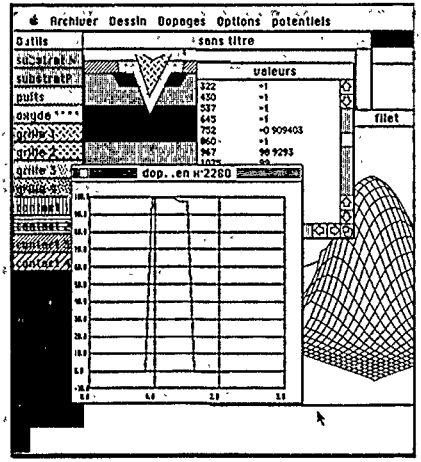


figure 1 : The computer screen

[1]This software has been developed in Think C™ version 4 which supports objetcs oriented programming (close to C++).

PROCESS SIMULATION

The software scans the structure to find zones concerned by the process simulation: pure doping implants and contacts[2]. A detailed 2-D profile has to be evaluated Mainly for implants. It is done through a diffusion process based on the numerical resolution of the *heat differential equation* on a rectangular mesh. For each zone, diffusion parameter values such as temperature, time and coefficients, are proposed to the user (by "dialogs") and can be adjusted. The software is able to build a finite difference system of equations which models the diffusion of the dopant in the local area around the implant. This depends of course on the local environment : oxide, substrate or boundaries... Several algorithms have been tested (alternate direction iteration), incomplete factorization on a Crank-Nicholson scheme in 2-D (Buleev-Stone[4]) appears to be the most efficient, especially when the two dimensions (vertical and horizontal) largely differ.

The evolution of the diffusion is also shown through clouds of dots evolving on the screen during the computation. The user can also look for - a concentration cross-section , - a list of values - and/or a critical value at a mouse-clicked point.

DEVICE SIMULATION

A rectangular global mesh is built. The addition of all the local distributions of dopants gives the global distribution When this is done, a new mesh is computed based on the variations from point to point, and interpolated values are evaluated. The procedure of re-meshing can also be launched manually. To prepare the device simulation, a table of characteristics of each point is built : The characteristics specify the nature of the equations corresponding to each point : known voltages, boundary with zero electric field, oxide (zones without charges)...

A new finite difference scheme is automatically built, indexed on the geographical changes in the device . POISSON's equation and continuity equations are solved. Coefficients for the corresponding difference equations are computed. Several discretization schemes for the current have been tested to tackle now well-known problems of consistency [1],[2],[3],[5] involvedby the equation of the current(MOCK's stream functions [6], MOTTET's intermediate continuity[7]). Taking advantage of memory available even now on small computers, large (and sparse) systems are built and solved by factorization in a sequence of iterations according to the NEWTON-RAPHSON algorithm. Here, the GUMMEL[5] iteration (decoupled equations) is used. Each simulation of an electrostatic state starts from an approximative equilibrium (to begin), or from a preceding state as an initial condition, or . The computation time depends naturally on the number of points of the mesh and the voltage values given to the simulator.

EXPLOITATION

The user is asked to give potentials for the contacts and/or gates. The simulator then solves the equations to evaluate the distribution of potentials and charges, from which are derived the currents.Actually only static behavior and relatively simple forms of equation are taken into account (Boltzmann statistics, constant mobilities), but the structure of the software should allow easy evolution.

Numerical simulations often produces an impressive quantity of data from which it is difficult to extract what is relevant to the user's problem. Our approach offers a great interactivity at this stage. The results (electrostatic potentials, charges ...) are accessible either using menus and/or through mouse-clicks on appropriate representation : net representation, lists of values in several *MacIntosh windows* .For instance, a click on equi-potential curves produces a new window with a cross section at the corresponding location. Results are also easily transferable to typical word processors spreadsheets or other traditional productivity softwares.

CONCLUSION

This software tool because of its intuitive access and direct connections with desktop work may be used in a great variety of situations: besides teaching and demonstrations of device behavior, the designers can "measure" the influence of a parameter, look at some bidimensional effects, without having to spend hours on learning how to use the tool. The structure of the software allows customization to take into account specific aspects. Some versions of this software[8] have been used to study aspects of noise in MOS transistors, as well as antiblooming designing in CCD cells[9].

REFERENCES

[1] SELBERHERR S. Analysis and simulation of semiconductor devices, Springer-Verlag ,Wien (Austria) 1984

[2] COLE D.C. and al. Solid-State Electron., 33, p 591, 1990.

[3] W.FICHTNER, D.J.ROSE. R.E.BANK, I.E.E.E. Trans on Electron.Dev. ED-30, 9, p.1018 (1983)

[4] H.L.STONE, SIAM J. Numer.anal. vol.5, n°3,sep 1968.

[5] D.SCHARFETTER, K.GUMMEL, I.E.E.E. Trans. Electron. Dev., ED 16, p.64 (1969)

[6] M.S.MOCK, Solid State Electron.,16, p.601 (1973)

[7] J.E.VIALLET, S.MOTTET Proc. of NaseCode IV Conf. Boole Press, Dublin, p.530 (1985).

[8] Y.MOREAU - SEE club modélisation, p 23, Paris (1990)

[9] Y.MOREAU - D. SELIMI submitted for publication

[2] Contacts are considered as doped zone : they might be N+ or P+ Silicon, and this renders band curvatures.

CIRCUIT SIMULATORS EMPLOYING TWO-DIMENSIONAL PHYSICAL MODELS OF POWER SEMICONDUCTOR DEVICES

Masa-aki Fukase*, Tadao Nakamura** and Jun-ichi Nishizawa***

* Research Institute of Electrical Communication, Tohoku University, Sendai 980, JAPAN
** Dept. of Machine Intelligence and Systems Engineering, Faculty of Engineering, Tohoku University, Sendai 980, JAPAN
*** President of Tohoku University, Sendai 980, JAPAN

Abstract-We present an accurate circuit simulator for power semiconductor devices by using 2-D physical device models and hierarchical p[r]ocessing of system equations based on Newton-Raphson's method. Performing the SI thyristor, we not only use the circuit simulator for analyzing its turn-off process but also design its superior cross-sectional structure. Finally we describe schemes for accelerating the hierarchical processing in terms of computer architecture.

I. INTRODUCTION

As for the DC analysis of semiconductor devices, it is satisfactory to compute only device models [1]. However, transient analysis requires to compute system equations comprising circuit equations and device equations. Some researchers have used the Gummel's algorithm [2] that simply iterates to calculate both equations as shown in Fig. 1 (a) [3], [4]. However, the Gummel loop does not always theoretically ensure the convergence of iterative solution. Then, hierarchical processing has been presented for system equations as shown in Fig. 1 (b) by employing Newton-Raphson's method to circuit equations that implicitly contain one-dimensional (1-D) physical device equations [5].

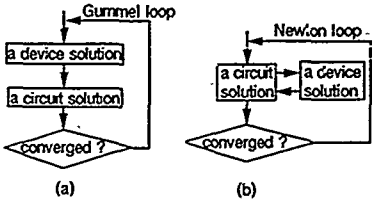


Fig. 1 Processing of system equations. (a) Gummel iteration; (b) Newton iteration.

We use the hierarchical processing for a circuit simulator of power semiconductor devices that should be approximated by 2-D physical models. 2-D physical device models located in a lower hierarchy are quickly and accurately calculated to determine the currents in circuit models. Thus, we realize the circuit simulator that is physically as well as mathematically accurate. Then, we use the circuit simulator for analyzing the turn-off process of a basic circuit containing the static induction (SI) thyristor.

II. ALGORITHMS OF THE HIERARCHICAL PROCESSING

Newton-Raphson's method yields for circuit equations

$$\begin{pmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} \end{pmatrix} \begin{pmatrix} \delta v_1 \\ \delta v_2 \end{pmatrix} = - \begin{pmatrix} f_1(v_1, v_2, i_1, i_2) \\ f_2(v_1, v_2, i_1, i_2) \end{pmatrix} \qquad (1)$$

where f_i, v_i, i_i, (i=1,2) represent a function of a circuit equation, a terminal voltage, and a terminal current, respectively. Currents and their partial differentials that are needed for processing Eq. (1) are computed from terminal currents of 2-D physical models. Therefore, the processing of Eq. (1) is hierarchical as shown in Fig. 2, where top-down traversal is done whenever a trial circuit solution is given in processing Eq. (1).

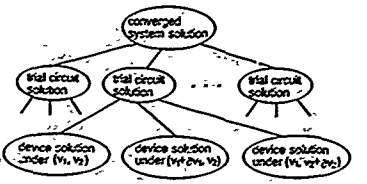


Fig. 2 Data structure formed in processing Newton iteration.

We use again Newton-Raphson's method for 2-D physical model equations in computing device solutions. The divergence of a solution in each of two nesting Newton loops is nipped in the bud, which is mathematically allowable. Then, a converged system solution is physically verified by examining whether the locus of an input/output working point is involved in a respective load surface.

In order to successfully obtain a converged system solution, it is important to quickly and accurately compute 2-D physical device models. The processor is occupied mostly by matrix equations whose computing time depends on both the number of mesh points and algorithms. In order to decrease the number of mesh points, we use an irregular mesh structure [6]. Moreover, we define an x/y differential mesh that is convenient to the Scharfetter-Gummel's integral [7], though a first order Taylor series is alternatively used in the place where the electric field intensity is nearly equal to zero. As for algorithms, CG methods recently are used more popularly [8], [9]. Nevertheless, it has not yet been discussed completely that CG methods are more effective than SOR methods [10]. Hence, we employ the SLOR method [11]. The simplicity of the SLOR method is preserved, though coefficient matrices are rather complex due to both the irregularity of the mesh structure and an extraordinary boundary condition assumed at an internal gate center.

Terminal currents are accurately calculated by counting displacement currents. Within a region of a device, the sum of a main component of conduction currents and that of displacement currents takes a constant value, which can be considered as a terminal current. Moreover, we partly make use of the Kirchhoff's law that mathematically stands among the terminal currents, because recombination/generation rates of electrons and holes are assumed to be equal each other.

Time for computing full periods of transient operation is strongly concerned with time increments Δt's. As for computing device solutions, Δt's are not restricted by any condition, because we have used the backward Euler method that is unconditionally stable, though large Δt's make it difficult to determine trial device solutions. Moreover, Eq. (1) does not explicitly contain Δt. However, large Δt's tend to make circuit solutions divergent because terminal currents vary drastically by the small change of terminal voltages. Consequently, Δt's are determined in this paper so that the variation of a main terminal current during $(t-\Delta t, t)$ is almost constant, which is suitable for converging both Newton loops.

III. TURN-OFF PROCESS OF THE SI THYRISTOR

Device analysts have not paid much attention to the crystal orientation of wafers to which process designers have attached great importance. Outside the ohmic region, drift velocities exhibit anisotropic behavior with respect to the orientation of an electric field applied to silicon crystals in common environment [12], which

naturally reminds us of the plane index dependency of device characteristics. Therefore, we basically view carrier mobilities as 2-D rank tensors [13]. Moreover, tensor elements are assumed to have separate relationships for electric field, temperature, and impurity density, respectively.

We use the circuit simulator for simulating the turn-off process of a basic circuit containing the SI thyristor [14]. Computed results are physically as well as mathematically self-consistent judging from Fig. 3. It is made clear from the simulation that the SI thyristor is hard to cause current crowding phenomena in turn-off states. On the other hand, the SI thyristor may have long tail periods.

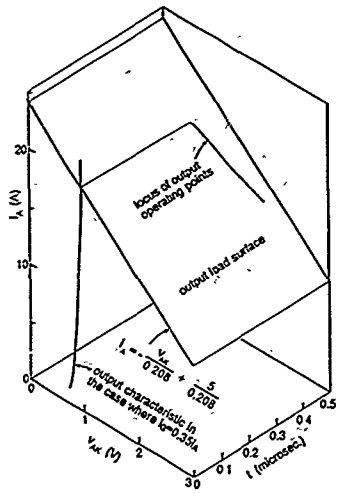


Fig. 3 Locus of an output operating point.

This problem has generally been avoided by composing a shorted anode emitter or consciously doping lifetime killers in a base. However, the former lacks reverse voltage blocking capabilities, and the latter causes large forward voltage drops. Thinning bases is a fundamental method to decrease stored holes or shorten tail periods, which is also useful for making forward voltage drops low. The cross sectional structure shown in Fig. 4 with larger impurity density of both gate center and anode surface, smaller impurity density of n^- base, and shorter channel width should have shorter t_{off}, higher forward/reverse voltage blocking capabilities, and lower forward voltage drops [15]. The n base layer added on the anode is useful for not only avoiding punch through phenomena but also shortening hole/electron lifetime in low injection level owing to their impurity dependencies, which is reasonably neglected in this paper. It is interesting to investigate the optimum wave form of the gate source voltage for quick turn-off operations.

IV. CONCLUDING REMARKS

Many issues have been made apparent. A physical quantity is defined in either differential meshes or a potential mesh. When one of the two mesh types is used in a finite difference formula containing both groups of quantities, the quantities belonging to the other mesh type should be modified, which neglects the definition of the mesh types. The distribution of space charge density may be hard to converge in iterative solutions as well as forward differences [16], which is not critical for the function of circuit simulators. Avalanche multiplication phenomena should not have been omitted.

We will consider two ways to accelerate the processing of Eq.

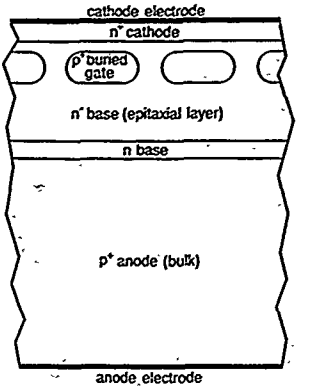


Fig. 4 Superior cross sectional structure of the SI thyristor.

(1) by viewing the tree shown in Fig. 2 in terms of computer architecture. One is to process in parallel three leaves that belong to the node of a trial circuit solution by using multiprocessors, though the tree traversal has been limited to serial processing in this paper. The other is to traverse the tree from its bottom to its top, which is realized by building up database of device solutions in advance. Moreover, we will determine as long Δt's as possible to quickly compute full periods of transient operations by investigating the relation between two Newton loops and the change of a main current.

REFERENCES

[1] M. Fukase and J. Nishizawa, 6th International Conf. on Control Systems and Computer Science, 3, pp. 106-118, May 1985.

[2] H. K. Gummel, IEEE Trans. on ED., ED-11, 10, pp. 455-465, Oct. 1964.

[3] L. J. Turgeon and D. H. Navon, ibid., ED-25, 7, pp. 837-843, 1978.

[4] A. Nakagawa, Solid-State Electronics, 28, 7, pp. 677-687, 1985.

[5] M. Kurata, R. Katoh, and J. Yoshida, ibid, ED-32, 6, pp. 1086-1091, June 1985.

[6] M. S. Adler, NASECODE I, pp. 3-30, June 1979.

[7] D. L. Scharfetter and H. K. Gummel, IEEE Trans. on ED., ED-16, 1, pp. 64-77, 1969.

[8] S. Nakamura and A. Nakagawa, Trans. of IEE Japan, 108, 5, pp. 333-338, 1988.

[9] A. Yoshii, H. Kitazawa, M. Tomizawa, S. Horiguchi and T Sudo, IEEE Trans. on ED., ED-29, 2, pp. 184-189, 1982.

[10] A. Yoshii, M Tomizawa and K Yokoyama, Solid-State Electronics, 30, 8, pp. 813-820, 1987.

[11] R S. Varga, Matrix iterative analysis, Prentice-Hall, Inc., 1962.

[12] C. Jacoboni, C. Canali, G. Ottaviani and A. A. Quaranta, Solid-State Electronics, 20, pp. 77-89, 1977.

[13] W. R. Runyan, Semiconductor measurements and instrumentation, McGraw-Hill, Ltd., 1975.

[14] M Fukase, T Nakamura, and J. Nishizawa, to be published by Trans. of IEE Japan.

[15] M Fukase, T Nakamura and J Nishizawa, contributed to IEEE Trans. on Power Electronics.

[16] M Fukase and J. Nishizawa, Trans of the IECE of Japan, E61, 4, PP. 329-330, April 1978.

IMPROVED COMPUTATION OF COEFFICIENTS FOR FEM IN ELECTRON OPTICS [1]

B. Lencová
Particle Optics Group,
Department of Applied Physics
Delft University of Technology
Lorentzweg 1
2628 CJ Delft, The Netherlands

and

M. Lenc
Institute of Scientific Instruments
Czechoslovak Academy of Sciences
Královopolská 147
CS-61264 Brno
Czechoslovakia

Abstract: The FEM with linear shape function does not provide a correct expression for the coefficients of FEM equations at axial points for the rotational symmetric and multipole scalar potentials. A correction of the coefficients for axial points is described.

1. Introduction

FEM is used for most computations of fields of electrostatic and magnetic lenses and deflectors in electron optics [1]. High accuracy of field computation is required e.g. for direct ray tracing of charged particles. We use the FEM with a topologically regular mesh of small quadrilaterals, subdivided further into triangles, where a linear shape function is used. The resulting matrix has a simple structure, and its solution is very fast with the preconditioned conjugate gradient method. However, in the two dimensional computations by FEM in cylindrical coordinates (r,z), there are some serious problems due to the use of the linear shape function in the evaluation of coefficients of FEM equations.

For rotational symmetric magnetic lenses, the discontinuity of the normal derivative of the vector potential $A(r,z)$ on the triangle boundary can give a large error e.g. in the case of high permeability cylindrical rods on the axis. The coefficients of the FEM equations depend on the way in which the integration of the term A/r in the energy functional is performed. We have shown that it is possible to find a method for the evaluation of coefficients which overcome this problem [2].

For the axially symmetric scalar potential problems the linear FEM provides the same coefficients as the five point finite difference method in rectangular meshes, except for the points on the axis [3]. In the computation of mth Fourier harmonics of the scalar potential Φ_m for deflectors and other multipoles Φ_m close to the axis is proportional to r^m, and thus the shape function linear in r does not provide correct results. This is possible to overcome by a formulation of the FEM equations for $\Psi_m=\Phi_m/r^m$ [4]. Even then the FEM equations for the points on the axis do not have the required accuracy. The paper presents a suitable method to correct for this behavior, valid both for the axially symmetric potential $\Phi_0(r,z)$ and for the function $\Psi_m(r,z)$.

2. The energy functional

In electron optics the axially symmetric lenses are mostly placed along the same axis. The electrostatic and magnetic deflectors and other multipoles are frequently defined on rotational symmetric formers, used in the presence of rotational symmetric materials and placed along the same electron optical axis. For some of the frequently used geometries of multipoles it is possible to evaluate the fields of these multipoles by making a decomposition into individual Fourier harmonics. For each mth harmonics the distribution of scalar potential $\Phi_m(r,z)$ can be evaluated as a two dimensional problem, with $\Phi_m=0$ on the axis and on the boundary far from the elements. In the vicinity of the z axis

$$\Phi_m(r,z)=r^m[d_m(z)-d_m''(z)r^2/4(m+1)+ +d_m''''(z)r^4/32(m+2)(m+1)- \ldots] \quad (1)$$

For $m=0$ is $\Phi_0(r,z)$ the axially symmetric scalar potential, which is not equal to zero on the axis. The potential $\Phi_m(r,z)$ is not suitable for the linear FEM if $m>1$ because of the dependence on r as r^m (1). Instead of Φ_m we define the FEM for the potential function $\Psi_m(r,z)=\Phi_m(r,z)/r^m$.

The radial derivative of the new function $\Psi_m(r,z)$ is close to the axis proportional to the distance from the axis and it is equal to zero on the axis. The linear shape function overestimates the effect of the radial derivative, and the field is next to the axis incorrect, e.g. for applications like the direct ray tracing. The consequence of this is, that even for regular rectangular mesh the coefficients of the FEM equations have lower accuracy than the five point finite difference formula [3]. For linear FEM, popular in electron optics, the mesh lines in the paraxial region are perpendicular to the axis.

Given the form of the potential function Ψ_m near the axis, a straightforward solution might be the use of r^2 as a new variable [5]. The way in which the finite difference formula for the axial points is derived is also based on this approach [6].

The energy functional for the mth component Φ_m contains the sum of squares of the r and z derivatives. For Ψ_m the energy functional is

$$E=c\iint\left\{\left[\left(\frac{\partial\Psi_m}{\partial r}\right)^2+\left(\frac{\partial\Psi_m}{\partial z}\right)^2\right]r^{2m+1}+\frac{\partial}{\partial r}\left(mr^{2m}\Psi_m^2\right)\right\}drdz \quad (2)$$

with $c=(\pi/2)(1+\delta_{m0})\mu$ for magnetic problems, for electrostatic ones is permeability μ replaced with the dielectric constant ε, δ_{m0} is the Kronecker δ. The first part of this energy functional (2) can be expressed in terms of shape function linear in r and z coordinates in a triangular finite element as

$$U_1=\frac{c}{D_1}\frac{R_{2m+1}}{(2m+2)(2m+3)}\sum_i\sum_j(\xi b_ib_j+c_ic_j)\psi_i\psi_j, \quad (3)$$

where $D_1=z_1(r_2-r_3)+z_2(r_3-r_1)+z_3(r_1-r_2)$ is twice the area of the triangle, $i,j=1,2,3$, ψ_i are the values of $\Psi_m(r,z)$ at the vertices, $b_1=r_2-r_3$, $c_1=z_3-z_2$; b_i, c_i for $i=2,3$ are obtained by cyclic interchange. ξ is a penalty factor equal to 1 everywhere except in the triangles with two vertices on the axis. This local penalty factor will be used later to correct for the too large effect of the radial term. R_n is an expression in terms of coordinates of the vertex

$$R_n=\sum_{k=0}^{n}\sum_{l=0}^{n-k}r_1^kr_2^lr_3^{n-k-l}.$$

Another possible expression can be derived by using a shape function linear in $u=r^2$ and z, again without the last term in (2), as

$$U_2=\frac{c}{D_2}\sum_i\sum_j\left[\frac{2U_{m+1}}{(m+2)(m+3)}c_ic_j+\frac{U_m}{2(m+1)(m+2)}d_id_j\right]\psi_i\psi_j \quad (4)$$

with $D_2=z_1(u_2-u_3)+z_2(u_3-u_1)+z_3(u_1-u_2)$, $d_1=u_2-u_3$, d_2 and d_3 cyclically. U_n is given by the same expression as R_n before, only using u instead of r.

[1] The work was performed as a part of FOM project IOP-IC-DTN 45.006.

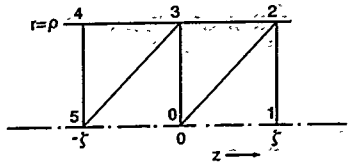


Fig. 1. A part of the rectangular mesh next to the axis and its subdivision into triangles.

Let us take a part of the fine mesh next to the axis (fig. 1). For simplicity we shall consider a rectangular mesh with a step ρ in the radial direction and a step ζ in the z direction. The FEM equations are obtained by the variation of the sum of terms in (3) for all triangles next to the central point with respect to potential ψ_0 there as

$$\psi_0\left[(2m+3)\,\zeta^2+\,2\xi\rho^2\right] = \left[(2m+3)\zeta^2\right]\psi_3+\,\xi\rho^2\left(\psi_5\,+\,\psi_1\right) \qquad (5)$$

or by variation using U_2 as

$$\psi_0\left[4(m+1)\,\zeta^2+\,2\rho^2\right] = \left[4(m+1)\zeta^2\right]\psi_3+\,\rho^2\left(\psi_5\,+\,\psi_1\right) \qquad (6)$$

If we substitute the expression (1) for ψ_m near the z axis into (5), we get

$$\psi_0=d_0 + \frac{\rho^2\zeta^2}{(2m+3)\zeta^2+2\xi\rho^2}\left[\left(\xi-\frac{2m+3}{4(m+1)}\right)d_0'' + \left[\frac{\xi\zeta^2}{12}+\frac{(2m+3)\rho^2}{32(m+1)(m+2)}\right]d_0''''\right] \qquad (7)$$

and similarly substituting into (6) we get

$$\psi_0=d_0 + \frac{\rho^2\zeta^2}{4(m+1)\zeta^2+2\rho^2}\left(\frac{\zeta^2}{12}+\frac{\rho^2}{8(m+2)}\right)d_0'''' \qquad (8)$$

We can see that the error in (7) is of the second order in the mesh size while the error in (8) is only of the fourth order.

If we now choose the factor $\xi=(1/4)(2m+3)/(m+1)$ for the triangles with two vertices on the axis, and use $\xi=1$ for all other triangles, we get the accuracy of the fourth order also for the points on the axis, without disturbing all the other features of FEM using shape functions linear in r and z coordinates.

In case of steps ζ different on the left and right hand sides of the point ψ_0 somewhat more complicated expressions are obtained. As expected, the first error term is proportional to the mean stepsize and difference in the stepsize, if only the subdivision shown in fig. 1 is used. Using also the other possible subdivision by a diagonal in the quadrilaterals [7], this error term vanishes, and the first error term improves by one order of magnitude in mesh step. It is therefore necessary to avoid large change of mesh density by using a mesh with smoothly varying stepsize.

3. An example

An example of the effect of the correction on the field computation will be given for an electrostatic axially symmetric potential for an arrangement of a simple geometry consisting of three electrodes. Fig. 2 gives the axial potential calculated in a rectangular mesh with different number of points within the radius, with and without the correction. The effect of the correction changes the axial potential in the opposite direction than the analytical solution is, due to the sign of the second derivative of the axial potential $\Phi_0(z)$. If no correction is applied, the axial field at a distance of one mesh line has an error of about 10 %. For multipoles a similar effect on the axial multipole function $d_m(z)$ can be obtained.

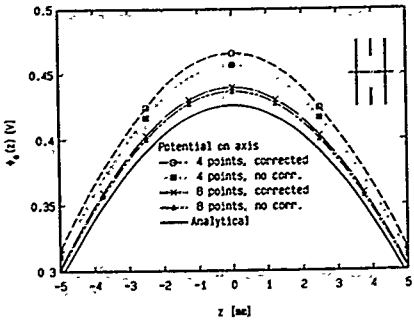


Fig. 2. Axial potential for the geometry shown in the corner (the distance between the outer electrodes and the inner electrode diameter being 20 mm). The outer electrodes are at zero potential, the central one has a unit potential. The correction to the maximum axial potential is about 1% for a coarse mesh with 4 points between the electrodes.

4. Programs

Fast and accurate field computations can be performed on PCs. Inside the standard 640 kbyte memory in DOS we can use meshes with up to 8400 mesh points [8]. The field computation programs are written in Fortran 77 and they run also on larger machines. They have a suitable user interface written in Turbo Pascal for IBM compatible PCs, allowing a graphical input and modification of data, automeshing, and they provide the input file for computations. They also allow the graphical display of outputs (e g. equipotentials, axial fields, etc) on screen or for hardcopy.

5. Conclusion

The advantage in using the improved coefficient computations are that we can use the FEM with linear shape function and simple mesh layout to obtain fast and accurate results. Very significant effect on the accuracy of computation brings the use of the multipole potential function Ψ_m, whereas the other corrections have only a slight effect on the axial field value. Although this effect as such may seem very small and visible only for small mesh sizes, it is an important improvement for the direct ray tracing, e.g with the slice method for obtaining fields from mesh potentials [9].

References

[1] E. Munro, these proceedings
[2] B. Lencová and M. Lenc, submitted to COMPUMAG 1991.
[3] E. Kasper and F. Lenz: Proc. 7th Eur. Congress on Electron Microscopy, (Leiden 1980), Vol. I, 10.
[4] B. Lencová, M. Lenc and K. D. van der Mast, J. Vac Sci. Technol. B7(19890),1846
[5] E. Munro, Proc. Intern. Symp. Electron Optics, Beijing 1986, 177.
[6] E. Durand, Électrostatique, Tome II Masson et cie, Paris 1966.
[7] B Lencová and M. Lenc, Scanning Electron Microsc 1986, Part III (SEM Inc., Chicago, 1986), 897
[8] B. Lencová and G. Wisselink, Nucl Instr Meth A298(1990),56.
[9] J. E. Barth, B. Lencová and G Wisselink, Nucl Instr. Meth. A298(1990),263.

THE NUMERICAL IMPLEMENTATION OF FOWLER-NORDHEIM TUNNELING AND INTERNAL FIELD EMISSION IN A GENERAL PURPOSE 2-D DEVICE SIMULATOR

Stephen Keeney[‡], Alan Mathewson[‡], Massimo Morelli, Leo Ravazzi,
Roberto Bez and Claudio Lombardi
SGS-Thomson, Central R&D, Agrate Brianza, Milano, Italy.
[‡]NMRC, University College Cork, Ireland

Abstract–**Fowler-Nordheim tunneling[1] and internal field emission[2] models have been incorporated into the 2-D numerical device simulator HFIELDS[3]. These phenomena are especially important when modeling a range of devices including DRAM, non-volatile memories and MOSFETS. The numerical implementation of the equations, the assumptions made and their justification will be discussed as well as a comparison of measured and simulated device characteristics.**

I. Introduction

The scaling of semiconductor devices as the ULSI era approaches is becoming increasingly difficult. This is prompting the need for more advanced design tools to allow the accurate simulation of advanced semiconductor structures prior to fabrication. More specifically the incorporation of F-N tunneling and internal field emission models into HFIELDS allows the efficient design of EEPROM (Electrically Erasable Programmable Read Only Memory) and flash EEPROM memory cells (Fig.1) by simulating the transient writing characteristics[4][5].

II. Fowler-Nordheim Tunneling (FN)

The model of Lenzlinger and Snow has been used to implement this phenomenon where the tunneling current density J_{FN} can be expressed as;

$$J_{FN} = \frac{q^3 m}{8\pi h m^*_{ox} \phi} E^2 \exp\left(-\frac{8\pi\sqrt{2m^*_{ox}\phi^3}}{3qhE}\right) \qquad (1)$$

where the symbols have their usual meaning. This is a 1-D model integrated over a suitable surface to calculate the total tunneling current. The 2-D electric field values are chosen so that the calculation takes proper account of the fringing fields. The integrating surface normally lies on the Si/SiO_2 boundary where the local field value is used for the calculation. This should be accurate as long as there is no significant charge trapped in the oxide bulk, which is the case in a non-degraded device. Charges trapped at the interface do not effect the accuracy of the model. Furthermore, different coefficients are selected for the FN equation depending on whether the emitting surface is mono or polycrystaline silicon. In the case of electrons tunneling through the silicon dioxide into monocrystaline silicon, electron-hole pair generation occurs because of the large energy difference between the oxide and silicon conduction bands. The carrier generation rate is fixed at 1.8 per tunneling electron which is typical of that reported in the literature[6]. The model has been verified on a tunnel oxide capacitor where a comparison of measurement and simulation is shown (Fig 2). If a floating gate is present, as is the case of an EEPROM, then the charge boundary value is continually updated with time according to the F-N current flowing. The model has been tested on an EEPROM device (1.2μm technology) during programming where measured and simulated tunnel currents show good agreement (Fig.3).

III. Internal Field Emission

This is commonly called band-to-band tunneling (BBT) and occurs in regions of semiconductor where high electric fields and large band bending are present. The model used is similar to the well known 1-D model of Kane[7]. In BBT the carrier transition probability can be expressed as,

$$T = \exp\left(-\frac{\pi m_r^{\frac{1}{2}} \epsilon_g^{\frac{3}{2}}}{2\sqrt{2}e\hbar E}\right)\exp\left(-\frac{\sqrt{2}\pi m_r^{\frac{1}{2}} \epsilon_1 \epsilon_g^{\frac{1}{2}}}{e\hbar E}\right) \qquad (2)$$

In this implementation the local field is used to calculate the carrier transition probability which introduces some error when the spatial variation of the electric field is not a constant. This error is largest when the spatial derivative of the electric field is a maximum. However, since the most significant internal field emission in a gated diode structure occurs in a very localized depleted region in the structure (Fig.4), in general choosing the local field introduces only a small error. Furthermore the amount of band bending must be evaluated at each element in the device since electrons will not tunnel into the forbidden band. The results of the model are shown in Fig.5 where a gated diode structure has been simulated and a comparison with measured values shows good agreement using only a single fitting parameter. Also shown in Fig 6 are the simulated (BBT) substrate currents seen during the erasing of a flash EEPROM device, emphasizing the sensitivity to oxide thickness.

IV. Conclusions

The techniques used to implement the F-N and internal field emission equations have been outlined and the models have been verified by comparison with measured data. The applications of these new models in HFIELDS has been briefly described in terms of non-volatile memory simulation which demonstrates the power of the tool as an aid in advanced microelectronics device design.

V. Acknowledgments

This work was partly funded by the European Community ESPRIT project 2039:(APBB). The authors would also like to acknowledge the contribution to code development by Massimo Rudan.

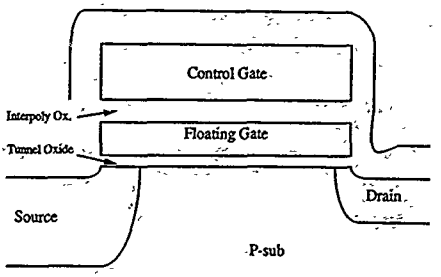


Fig.1. Cross-section schematic of a flash EEPROM device.

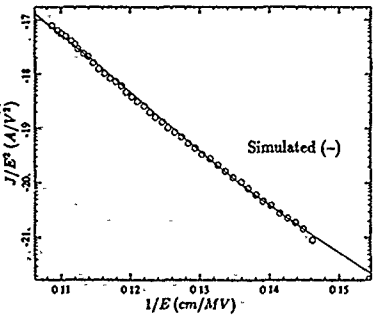


Fig.2. Fowler-Nordheim plot of a tunnel oxide capacitor.

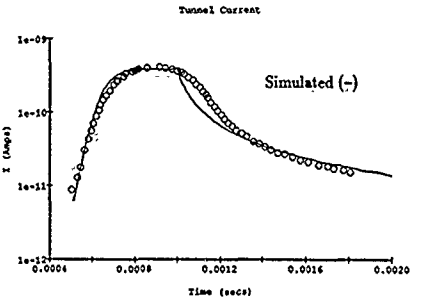


Fig.3. Simulated (–) and measured programming current in an EEPROM cell.

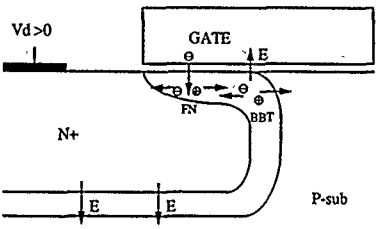


Fig.4. Cross-section of a gated diode showing the BBT and FN currents.

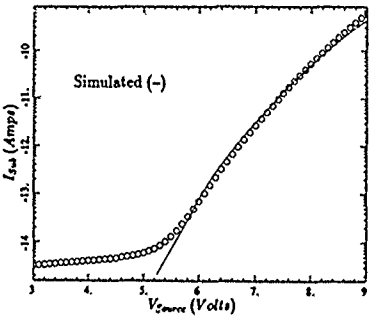


Fig.5. BBT current of a gated diode structure in reverse bias.

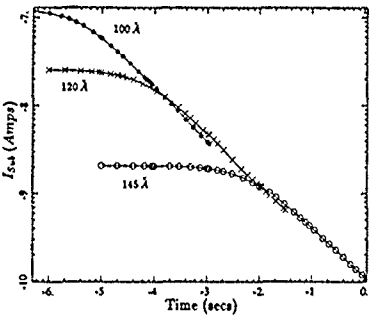


Fig.6. Simulated substrate current seen during the erasing of a flash EEPROM device.

References
[1] M.Lenzlinger et al.; *J Appl. Phys.*, vol.40 p278,1969.
[2] Z.A. Weinberg et al.; *J. Appl. Phys.*, vol.53.,p5082,1982.
[3] G.Baccarani et al.;*NASECODE IV, Tech. Dig.*, Dublin, Irl, Jun 1985
[4] S.Keeney et al. *IEDM90* p201, San Francisco, California,Dec 1990.
[5] S.Keeney et al. *IEEE NVMWS* Monterey, California,Feb. 1991.
[6] Kolodyne et al. *IEEE Trans Elec. Dev* ,vol. ED-33, p.835, Jun 1986.
[7] E O.Kane; *Phys. Rev.*, vol. 159 p264, 1967.

SIMULATION OF VERY HIGH SPEED BIPOLAR TRANSISTORS FABRICATED IN THE N-WELL OF A C-MOS PROCESS AT LOW TEMPERATURES

Chinmay K Maiti
Dept of Electronics & ECE
Indian Institute of Technology
KHARAGPUR 721302 India

ABSTRACT

Bipolar transistors designed for room-temperature operation suffer serious current gain degradation at low temperatures and have not been widely investigated for their operations and performances at low temperatures. In this paper, bipolar transistors fabricated in the n-well of a C-MOS process have been simulated using BIPOLE program at low temperatures. It is shown that at a temperature as low as 125°K, it is possible to obtain a cut-off frequency of 7 GHz and a current gain of 20 which is sufficient for some applications.

INTRODUCTION

Recent advances in high temperature superconductors [1] are giving new impetus to the understanding of bipolar and MOS devices operation and performances down to liquid nitrogen temperature. BICMOS has also received considerable attention as an attractive room temperature VLSI technology. Very high speed poly-emitter bipolar transistors fabricated in the n-well of a CMOS process, having room temperature cut-off frequency of about 16 GHz and an ECL gate delay time of 65 pSec have been reported [2,3]. Little attention has been paid to date regarding the performances of these devices at low temperatures. Numerical simulation has been used [4-6] as a tool to study the influence of different process parameters on bipolar device terminal behavior without going into a costly fabrication and measurement procedures.

The computer simulation results obtained using the BIPOLE device simulator have been validated first with the reported room temperature experimental results [2,3] in relation to the current gain, cut-off frequency and ECL gate delay time and then the results for low temperature simulations are presented.

STRUCTURE INVESTIGATED

The transistor studied in the present simulation is an integrated circuit device with a polysilicon emitter and uses oxide isolation. The emitter dimensions are 0.6 x 2.4 μm^2 and the e-b and c-b junction depths were 0.04 and 0.14 μm respectively. The extrinsic base sheet resistance was 3.0 Ω/sq and the collector was characterized by an epitaxial layer doping of 8×10^{17} cm^{-3}.

ANALYSIS

The simulations have been performed using the BIPOLE program [7] over the temperature range from 300° to 77° K. It should be noted that BIPOLE program includes physical effects such as heavy doping bandgap reduction [8], doping-level-dependent mobility, field-dependent mobility [9] and recombination rate versus doping level [10]. The program is also capable of identifying the contributions of charges in the various regions to overall delay time [11]. Special attention has been given to the device parameters such as current gain, cut-off frequency (fr) and ECL gate delay time. It is seen that the room temperature simulation results as described below, agree well with the experimental results reported in [2,3].

Figure 1 shows the variation of cut-off frequency (fr) as a function of collector current. The reported cut-off frequency is about 16 GHz (cf.figure 15 [3]) where as the simulated cut-off frequency is seen to be about 18 GHz.

Figure 2 shows the variation of current gain as a function of temperature. The reported room temperature current gain is about 100 (cf. Fig. 11 [3]) and simulated current gain is about 105. It may be seen that even at 125° K a gain of about 15-20 may be achieved for the device considered and is sufficient for some applications.

The experimental ECL gate delay time for 21 stage ring oscillator is about 65 pSec for fanout of 1 and tree current of around 0.4mA with a logic swing of 0.5V at CL=0.0 pF. The simulated delay time (computed using the expression reported in Ref.[12]) is about 75 pSec as shown in Figure 3.

Figure 4 shows the variation of peak cut-off frequency as a function of temperature and it is seen that even at 125° K, it is possible to have an fr of about 7 GHz. The variation of ECL gate delay time as a function of temperature is shown in Figure 5. It is interesting to note that around 0.3 mA collector current the minimum gate delay time of about 70 pSec is obtained.

CONCLUSIONS

The computer simulation results agree well with the experimental results reported and are sufficiently accurate to give confidence in interpreting the results. More experimental measurement data are necessary at low temperatures to confirm the temperature dependence of fr and current gain of bipolar transistors.

ACKNOWLEDGEMENTS

The author is grateful to Professor D J Roulston of the University of Waterloo, Canada for the BIPOLE program and wishes to thank Dr. B K Singh of the Computer Center for his help in computations.

REFERENCES

1. S.K. Tewksbury, L.A. Hornak and M. Hatamian:Solid State Electronics, vol.32, no.11,pp.947-959, November 1989.
2. T. Yamaguchi and T.H. Yuzuriha:IEEE Trans. on Electron Devices, vol.ED-36, no.5, pp.890-897, May 1989.
3. T. Yamaguchi et al.:IEEE Trans. on Electron Devices, vol.ED-35, no.8., pp.1247-1257, August 1988.
4. P. Ashburn and C.K. Maiti: Optimization of bipolar transistors for cut-off frequency, Tech.Report, Southampton Univ., April 1989.
5. M. Chrzanowska-Jeske and R.C. Jaeger: IEEE Trans. on Electron Devices,vol.ED-36,no.8, pp.1475-1488, August 1989.
6. S Selberherr: ibid.,pp.1464-1474.
7. D.J.Roulston: Proc. IEEE Custom Integrated Circuits Conf., Rochester, NY,May 1980,pp.2.
8. J.W. Slotboom: Solid State Electronics, vol.20, pp. 279-283, 1977.
9. N.D. Arora et al.:IEEE Trans. on Electron Devices, vol.ED-29, no.2, pp.292-295,Feb.1982
10. D.J. Roulston et al.: ibid., pp.284-291.
11. D.J. Roulston and F. Hebert:IEEE Trans. on Electron. Dev. Letts., vol. EDL-7, no.8, pp.461-462, August 1986.
12. E.-F. Chor et al.: IEEE Trans. on Solid State Circuits,vol.23,pp.251-259,Feb. 1988.

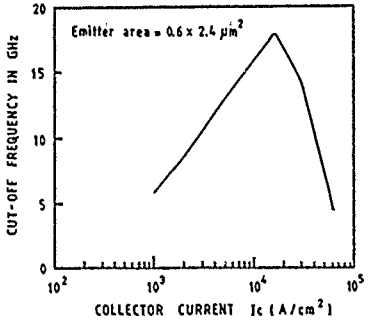


Fig. 1

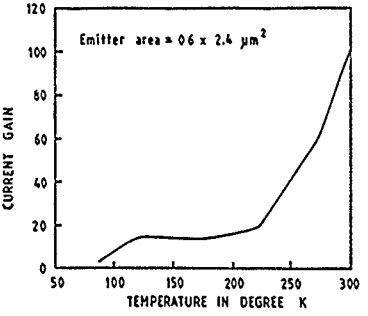


Fig. 2

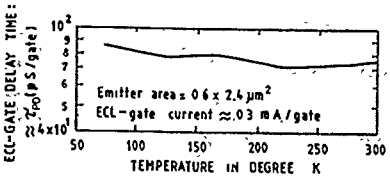


Fig. 3

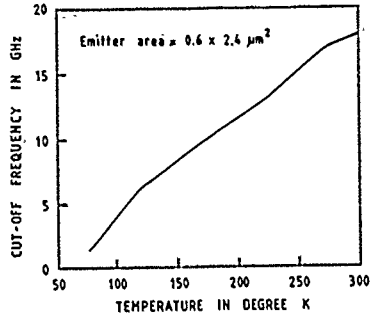


Fig 4

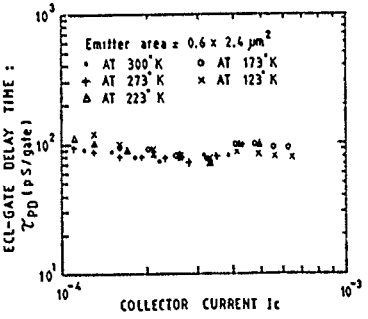


Fig. 5

HYBRID FINITE ELEMENT MODELS FOR 3-D SEMICONDUCTOR DEVICE SIMULATION

ANDERHEGGEN E.
KORVINK J. G.
SARTORIS G.
Swiss Federal Institute of Technology
ETH-Hönggerberg, HIL F23.3
CH-8093 Zurich, Switzerland

ROOS M.
SCHWARZENBACH H.U.
Corporate Research & Development
Landis & Gyr Betriebs AG
CH-6301 Zug, Switzerland

UNGRICHT H.
Technikum Winterthur Ingenieurschule
CH-8401 Winterthur, Switzerland

Abstract: A 3-D hybrid finite element model for the stationary semiconductor equations is presented. Symmetry group theory is used for each of the three PDE's to derive seven weak current compatibility equations for an 8-noded trilinear brick element leading to a stable (i.e. avoiding spurious modes) hybrid element discretization. It is also shown that more than seven linear independent weak current compatibility equations can not be introduced.

I. INTRODUCTION.

Several authors have transferred the concept of "Hybrid Elements" or "Mixed Element Discretization" developed in computational mechanics [2],[3] to semiconductor device simulation [1], [4], [5], [6] during the last three years. This dimension-independent concept reduces, in the 1-D case, to the standard Scharfetter-Gummel current discretization scheme [9] widely used in semiconductor device simulation. Numerical experiments showed that 2-D hybrid elements lead, at least in some 2-D examples, to more accurate results than the conventional box-method in conjunction with the Scharfetter-Gummel current approximation [1]. Crucial in a hybrid element model for the semiconductor equations is the "correct" selection of the trial current- and weight vectors, and the corresponding additional equations in order to determine the inner degrees of freedom (DOF's) introduced for the current assumption [6]. A "bad" selection of trial vectors lead to rank deficiency of the element stiffness matrix, resulting in the phenomena of "spurious modes" [7].

Punch and Atluri [8] used group theory to derive trial functions for the stress tensor components (mechanics problems) leading to a stable 3-D hybrid element (i.e. no spurious kinematic modes) which is also optimal with respect to computation time.

The goal of our paper is to present a 3-D hybrid brick element model (section III) to solve the stationary semiconductor equations, i.e. the Poisson and the continuity equations, for the unknown scalar fields, namely the electric potential and the Slotboom variables (cf. model problem defined in section II). Coordinate invariance entails certain symmetry relations between the coordinates. These relations are governed by group theory. Group theory is used for each PDE to define seven weighting vectors with polynomial components of minimal order, corresponding to exactly seven weak current compatibility equations (wcce's) which guarantee a stable brick element discretization with an 8x8 stiffness matrix of rank seven. Eight trilinear shape functions corresponding to the eight corner DOF's are used in this case for the unknown scalar field. These weighting vectors guarantee zero current distribution for constant scalar field distribution (rigid-body-motion condition). Furthermore, these weighting vectors are divergence free: No Lagrangian multipliers need be introduced. The two following statements will also be proven: a.) Each introduction of additional wcce's with a new weighting vector linearly independent from the previously defined seven wcce's and compatible to the rigid-body motion condition is inconsistent, i.e. results in a condition for the inner DOF's in the current assumption only. This condition is independent from the global DOF's of the unknown scalar field. b.) Leaving out anyone of the seven specified wcce's leads to instability (spurious modes).

II. BOUNDARY VALUE MODEL PROBLEM.

We consider the boundary value problem

$$\operatorname{div} \underline{E} = \rho \text{ in } \Omega \tag{1}$$

for an unknown C_0-continuous scalar field f, related to the vector field $\underline{E}$ by $\underline{E} = a(f, x,y,z)\,\mathbf{grad}f$. The scalar function $a(f, x,y,z)$ is as yet unspecified. The boundary $\partial\Omega$ of the 3-D simulation domain Ω is split up into two disjoint parts. Dirichlet boundary conditions are applied on the first part and homogeneous Neumann conditions ($\underline{E}\,\underline{\nu} = 0$, $\underline{\nu}$ being the outward unit normal vector to the boundary) on the second part. The inhomogeneous part $\rho(f, x,y,z)$ of (1) is a given function of f and the coordinates (x,y,z). The regularity conditions for $\underline{E}$ and ρ in order to guarantee existence and uniqueness of the solution of this problem are summarized e.g. in [9]. Each of the three stationary semiconductor device equations is of the form (1) using the Slotboom variables u and v (exponential functions of the scaled Fermi potentials) as independent variables. An extension of the following derivation for a set of other independent variables is possible [1]. It should be also observed that the derivation applies to both Poisson's and the continuity equations.

III. 3-D FE-HYBRID ELEMENTS FOR THE SEMICONDUCTOR EQUATIONS.

Following Piaa's hybrid model [2], an independent assumption for the vector $\underline{E}$ is introduced within each element volume Ω_e of a given mesh in the 3-D domain Ω

$$\underline{E}_a = \sum_{j=1}^{M} \beta_j \underline{E}_j(x,y,z) \quad \text{in } \Omega_e, \tag{2}$$

with M DOF's β_j.

Moreover, another C_0-continuous assumption is introduced within each element Ω_e for the scalar field f, defined by

$$f_a = \sum_{i=1}^{N} f_i H_i(x,y,z) \text{ in } \Omega_e, \tag{3}$$

with N nodal degrees of freedom f_i and with shape functions H_i. This assumption is C_0-continuous over the whole simulation domain Ω. However, compatibility between the vector field $\underline{E}_a$ and the scalar field f_a is not guaranteed within each element Ω_e.

According to the weighted residual method, two sets of discretized equations are considered, firstly the weak form of the vector field compatibility (wcce in the case of a continuity equation) for each element

$$\int_{\Omega_e} \underline{W}_j \; a^{-1} \, (E_a - E\,(f_a, f_{a_x}, f_{a_y})\,)\; d\Omega \;=\; 0 \quad \text{in } \Omega_e\,, \qquad (4)$$

$$(j = 1 \ldots P,\; P <= M)\,, \qquad (\text{assuming } a^{-1} = a^{-1}(f_a, x, y, z) \neq 0).$$

with weighting vectors $\underline{W}_j$, and secondly, the weighted conservation of flux of the vector field $\underline{E}$ over the whole domain Ω

$$\int_{\Omega} H_i \; (\mathrm{div} E_a - \rho\,)\; d\Omega \;=\; 0\,, \qquad (5)$$

with C_0-continuous weighting functions H_i corresponding to each DOF f_i. For a further derivation of element residual equations and generalized element stiffness matrix, see [1].

The question is: How are the $\underline{W}_j(x,y,z)$ and $\underline{B}_j(x,y,z)$ defined and what are the additional equations if P<M ?

IV. APPLICATION OF GROUP THEORY TO THE WEAK CURRENT COMPATIBILITY EQUATIONS.

Consider a cubic element with element faces parallel to the global coordinate axis. 8 trilinear element shape functions H_i are introduced corresponding to the 8 element corners, see Figure below.

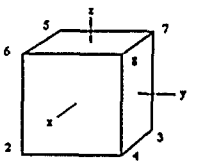


Figure. Element corner numbering

Writing the wcce's (4) in matrix form and applying one of Green's formulas to (4), cf. [1], the matrix

$$G_{ij} \;=\; \int_{\partial\Omega_e} \underline{W}_i \times H_j \; d\omega \qquad (6)$$

has to be evaluated. This matrix describes the coupling between the β_j in (2) and the f_i in (3).

The seven divergence-free weighting vectors are (local coordinate origin at element centroid):

$$\underline{W}_1 = [1, 0, 0]^T \quad \underline{W}_4 = [0, z, y]^T \quad \underline{W}_7 = [yz, zx, xy]^T$$

$$\underline{W}_2 = [0, 1, 0]^T \quad \underline{W}_5 = [z, 0, x]^T$$

$$\underline{W}_3 = [0, 0, 1]^T \quad \underline{W}_6 = [y, x, 0]^T \qquad (7)$$

Consider the eight symmetrized functions

$$h_0 = (H_1 + H_2 + H_3 + H_4 + H_5 + H_6 + H_7 + H_8)\; |\partial\Omega_e$$

$$h_1 = (-H_1 + H_2 + H_3 - H_4 + H_5 - H_6 - H_7 + H_8)\; |\partial\Omega_e$$

$$h_2 = (H_1 + H_2 - H_3 - H_4 - H_5 + H_6 - H_7 + H_8)\; |\partial\Omega_e$$

$$h_3 = (H_1 - H_2 + H_3 - H_4 - H_5 - H_6 + H_7 + H_8)\; |\partial\Omega_e$$

$$h_4 = (H_1 - H_2 - H_3 + H_4 + H_5 - H_6 - H_7 + H_8)\; |\partial\Omega_e$$

$$h_5 = (-H_1 + H_2 - H_3 + H_4 - H_5 - H_6 + H_7 + H_8)\; |\partial\Omega_e$$

$$h_6 = (-H_1 - H_2 + H_3 + H_4 - H_5 + H_6 - H_7 + H_8)\; |\partial\Omega_e$$

$$h_7 = (-H_1 - H_2 - H_3 - H_4 + H_5 + H_6 + H_7 + H_8)\; |\partial\Omega_e \qquad (8)$$

(the shape functions are restricted to the element boundary). Note that this is a regular transformation.

Result: The 7x8 matrix

$$T_{ij} = \int_{\partial\Omega_e} \underline{W}_i \times h_j \; d\omega\,, \qquad (9)$$

is an "upper-diagonal" matrix: $T_{ij} \neq 0$ for $i+1=j$ and $T_{ij} = 0$ for $i+1 \neq j$. It follows that the rank of T is equal to the rank of G and also to the rank of the stiffness matrix.

Introducing a new weighting vector $\underline{W}_I$, I>7, the row T_{Ij} can be described by linear combinations of rows of T_{ij}, i<=7 (a constant scalar distribution f_a over the element leading to zero current distribution implies $T_{I1} = 0$).

Theoretical background: The decomposition of matrix T is a direct consequence of the generalization of Schur's lemma [10]: Consider the symmetry group C consisting of all 48 rotations and reflections which leave the cube invariant. The linear maping

$$G:\; \underline{W}_i \text{ ----------> } \underline{W}_i \times |\partial\Omega_e \qquad (10)$$

whose matrix form is defined by T, commutes with every other element symmetry transformation. As a consequence, irreducible representation spaces of C contain the eigenspaces of G.

Thus the $\underline{W}_j$ and h_i transform according irreducible representations of C [10]. Equations (8) are the projection operators, well known in representation theory.

V. CONCLUSION.

Using symmetry relations, we showed that at least seven wcce's are needed to avoid spurious modes for a cubic trilinear element. A simular derivation of wcce's is possible if quadratic and cubic or other shape functions H_i are introduced. An extension of the theory to brick elements is possible. It is hoped that the selected weig'.ts are "adequate" for moderately distorted brick elements, compare [11].

In the case of a = constant, the same weighting vectors can be used for the vector field assumption $\underline{E}_a$ (Poisson's equation). In this case, the first term in equation (4) splits in blocks along the diagonal, cf. [10], resulting in some computational speed up.

If a ≠ constant, at least seven vectors have to be introduced in the vector field assumption (continuity equations).

If more than seven trial current vectors are used in $\underline{E}_a$, additional equations have to be defined (for example integrability conditions for $a^{-1}\underline{E}_a$ or inhomogeneity conditions for $\mathrm{div}\underline{E}_a = \rho$).

Note that in the case of $\underline{B}_j = \underline{W}_j$ for j = 1, ..., 7 leads to symmetric element stiffness matrices.

REFERENCES.

[1]: Anderheggen, E., et. al., "Numerical Comparison of Pian's 2-D Hybrid FE-Model With Some Classical device Simulation Discretization Methods Using Sever's Test Diode", NASECODE VI Conf. Proc., pp 441-447, Boole Press Dublin, 1989.

[2]: Pian, T.H.H., "A historical note about 'hybrid elements' ", Short comm., 1977.

[3]: Pian, T.H.H., "Derivation of Element Stiffness Matrices by Assumed Stress Distribution", AIAA, Vol. 2, 1964.

[4]: Polak, S.J., et. al , "A Finite Element Method With Current Conservation", SISDEP 88 Conf. Proc., pp 453-462, Tecnoprint Bologna, 1988.

[5]: Wang, S., et.al., "Mixed Finite Element Approximation of the Stationary Semiconductor Continuity Equations", SISDEP 88 Conf. Proc., pp 475-484, Tecnoprint Bologna, 1988.

[6]: Brezzi, F., et. al., "Mixed exponential fitting schemes for current continuity equations", NASECODE VI Conf. Proc., pp 546-555, Boole Press Dublin, 1989.

[7]: Pian, T.H.H., et al., "On the Suppression of Zero Energy Deformation Modes ", Int. J. Numer. Meth. Eng., 19, pp 1741-1752, 1983.

[8]: Punch, E.F., Atluri, S.N., "Applications of Isoparametric Three-Dimensional Hybrid-Stress Elements with Least-Order Stress Fields", Computers and Structures, Vol 19 No. 3, pp 409-439, Pergamon Press LTD., 1984.

[9]: Markowich, P.A , "The Stationary Semiconductor Device Equations", Springer Verlag, 1986.

[10]: Stiefel, E., et. al., "Gruppentheoretische Methoden und ihre Anwendungen", Teubner Verlag Stuttgart, 1979.

[11]: Punch, E.F., Atluri, S.N., "Development and Testing of Stable, Invariant, Isoparametric Curvilinear 2- and 3-D Hybrid-Stress Elements", Computer Methods in Applied Mechanics and Engineering, Vol 47, pp 331-356, North-Holland, 1984.

ON AN INTEGRAL EQUATION DESCRIBING THE SWITCHING BEHAVIOUR OF PN-DIODES

CHRISTIAN SCHMEISER
Inst. für Angewandte und Numerische Mathematik
TU Wien
Wiedner Hauptstr. 8-10
A-1040 Wien

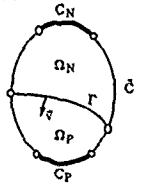
ANDREAS UNTERREITER
Fachbereich 3 der TU Berlin
Straße des 17. Juni 136
D-1000 Berlin 10

Abstract

The standard model describing the switching behaviour of a PN-diode consists of the well-known Drift-Diffusion-Equations of van Roosbroeck. Under simplifying assumptions (zero space charge approximation, low injection limit) an integral equation just in terms of the time-dependent current I(t) can be obtained from these coupled partial differential equations. A singular perturbation analysis of the integral equation yields a mathematical justification of the physically observable constant-current phase.

I. INTRODUCTION

The investigated physical system is an electrical circuit that consists of a voltage-source generating a potential drop U(t), a serial Ohmic resistance ω and a PN-diode. U(t) is assumed to be piecewise constant for $t > 0$ and $t < 0$, respectively. At $t = 0$, U(t) changes abruptly from +W (biasing in forward direction) to -U (reverse biasing). The physically most interesting quantity describing this abrupt switching is the time-dependent current I(t). Via identification of the PN-diode with a bounded domain $\Omega \subset \mathbb{R}^n$, n=1,2,3, see Fig.1, the standard (semi-classical) model describing the diode's



$\Omega = \Omega_N \cup \Omega_P$

Fig.1.

dynamics given by the (scaled) Drift-Diffusion-Equations of van ROOSBROECK [vanROOSBROECK,1950] is applicable. (For the scalings see [MARK,RING,SCH,1990].)

$$\lambda^2 \Delta V = n - p - C \tag{D1}$$
$$J_n = \mu_n(\nabla n - n\nabla V) \tag{D2}$$
$$J_p = -\mu_p(\nabla p + p\nabla V) \tag{D3}$$
$$\mathrm{div} J_n = \partial_t n + R \tag{D4}$$
$$\mathrm{div} J_p = -\partial_t p - R \tag{D5}$$

The physical quantities are the electric potential V, the (non-negative) concentrations of electrons (n) and holes (p), the doping profile C (which is strictly positive/negative in the N-region/P-region Ω_N / Ω_P) and the electron and hole current densities J_n and J_p. μ_n and μ_p denote the electron and hole mobilities, respectively, which are assumed to be functions of the spatial variable x taking positvie values. The differential equations involved are, respectively, the Poisson equation for the electric potential V, the continuity equations for electrons and holes and the electron and hole current relations. The spatial variable ranges over Ω, the time variable t ranges over $(0,\infty)$. R denotes the Shockley-Read-Hall-recombination term

$$R = \frac{np - \delta^4}{\tau_n(p+\delta^2) + \tau_p(n+\delta^2)} \tag{R}$$

based on mass-action-models of the electron-hole-interaction. τ_n and τ_p are the mean life-times of electrons and holes. They are assumed to be constant in the N-region and the P-region, respectively. Equations (D1)-(D5) are subjected to the boundary conditions (B1) along C_N, C_P:

$$\begin{aligned} &n - p - c = 0 \\ &np = \delta^4 \\ &V = V_{bi}(x) + U + \omega I(t) \quad , \ x \in C_N \\ &V = V_{bi}(x) \quad , \ x \in C_P \end{aligned} \tag{B1}$$

where

$$I(t) = \int_\Gamma (J_n + J_p - \lambda^2 \partial_t(\nabla V)) \cdot \vec{\nu} \, d\sigma \tag{I}$$

$$V_{bi}(x) = \ln\left[\frac{C + \sqrt{C^2 + \delta^4}}{2\delta^2}\right] \tag{V}$$

to homogeneous vonNeumann conditions along $\tilde{C}$.

$$(\nabla n)\cdot\vec{e} = (\nabla p)\cdot\vec{e} = (\nabla V)\cdot\vec{e} = 0 \tag{B2}$$

and to initial conditions:

$$n(.,0) = n^I(.), \ p(.,0) = p^I(.), \ V(.,0) = V^I(.) \tag{D^I}$$

where n^I, p^I, V^I are solutions of the stationary Drift-Diffusion-Equations. Under simplifying - but physically reasonable - additional assumptions (smoothness of Ω and the physical parameters, zero space charge approximation, low injection limit) a nonlinear Volterra Integral Equation of the second kind with I(t) as unknown function can be obtained (see [UNTERREITER,1991], [SCH, UNT, WEISS, 1991]):

$$e^{(-U-\omega I(t))} = 1 + cI_F + ((I - I_F) * g)(t) \tag{1a}$$
$$e^{(W-\omega I_F)} = 1 + cI_F \tag{1b}$$

I_F denotes the (scaled, dimensionless) constant current that has flown through the device for $t<0$. c cumulates physical parameters (such as the mean life-time of electrons and holes) of the diode. '*' denotes the convolution operation. The integral kernel g is an infinite sum:

$$g(t) = \sum_{j=1}^{\infty} (K_j)^2 \exp(\lambda_j t) \tag{2}$$

g is a completely monotone $L^1(0,\infty)$-function with

$$\lim_{t\to\infty} (1 * g)(t) = c \tag{3}$$

The set $\{(\lambda_j, K_j)\}$ is determined by the eigenvalue-problem

$$L[\phi_j] = \lambda_j \phi_j \tag{4a}$$

where ($\gamma = |C|$)

$$dom(L) \subset H_0^1 = \{f \in H^1(\Omega) : f|_{C_\alpha} = 0, \alpha \in \{N,P\}, \sqrt{\gamma} f|_\Gamma = \text{const}\} \tag{4b}$$

$$L_{H^2}[u] = \sqrt{\gamma}\,\mathrm{div}\left(\frac{\mu}{\gamma}\nabla(\sqrt{\gamma}u)\right) - \frac{u}{\tau} \tag{4c}$$

(extended weakly to dom(L))

and

$$K_j = \sqrt{\gamma \phi_j} l_P \qquad (4d)$$

μ and τ are the mobilities and mean lifetimes of the minonty charge carriers in the N- and P-region of the diode.
Furthermore, $0 > \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots$ and

$$\lim_{j \to \infty} \lambda_j = -\infty \qquad (4e)$$

(See [UNTERREITER,1991] for further details.)

II. THE EXISTENCE OF A CONSTANT-CURRENT PHASE (*)

One of the most significant features of the physically observable current I(t) is its time-independence for a certain period [0,T]. The justification of the existence of such a "constant-current phase" as well as the calculation of its duration can both be facilitated by a singular perturbation analysis of the integral equation (1a).
The scaling factor of the potential U in (1a) is given by the so-called "thermal voltage" U_T which is typically O(0 025Volt). U_T is rather small in comparison with the unscaled versions of U, $W_D = W \cdot \omega l_F = O(1\,\text{Volt})$. Consequently, it is advisable to introduce

$$\varepsilon = \frac{1}{U + W_D} \approx 10^{-2} \qquad (5)$$

with U and W_D being scaled. On the other hand, the derivation of (1a) essentially depends on the assumption that I(t) is well-scaled during the whole switching process (low injection). Therefore,

$$\eta = \frac{I(0+)}{I_F} \qquad (6)$$

is a well-scaled quantity. Introducing $z(t) := \omega I(t)$ and

$$\rho = \frac{\eta \omega}{c}$$

(1a) becomes in terms of z(t)

$$\exp(z - \frac{1}{\varepsilon}) = 1 - \frac{1}{c(1+\varepsilon\rho)} * g - \frac{\varepsilon\eta}{c(1+\varepsilon\rho)} z * g \qquad (7)$$

Referring to [MILLER,UNTERREITER,1991] and [SCH, UNT, WEISS, 1991],

$$\frac{1}{\varepsilon} = z(0) \geq z(t) \geq z_\infty = \lim_{t \to \infty} z(t) \qquad (8)$$

$z(0) = (1/\varepsilon)$ suggests the ansatz $z = (1/\varepsilon) + y$ for sufficiently small ε. (7) becomes in terms of y

$$e^y = 1 - \frac{1+\eta}{c(1+\varepsilon\rho)} * g - \frac{\varepsilon\eta}{c(1+\varepsilon\rho)} y * g \qquad (9)$$

As long as $y = o(1/\varepsilon)$ for $\varepsilon \to 0$ on an interval [0,T], y can approximatively be neglected and z is almost constant.

Theorem
Let $0 \leq \varepsilon < (\eta/\rho)$ and define $t_\varepsilon > 0$ such that

$$1 - \left[\frac{1+\eta}{c(1+\varepsilon\rho)}\right] * g(t_\varepsilon) = 0 \qquad (10)$$

1) If $y = o(1/\varepsilon)$ for $\varepsilon \to 0$ on [0,T], then $T \leq t_0$.

2) If $g(t) = O(t^{-\alpha})$ for $t \to 0$, $0 \leq \alpha < 1$, then $y = O(\ln\varepsilon) = o(1/\varepsilon)$ on $[0, t_0]$.

Part 2 of this Theorem guarantees the existence of a constant-current phase - provided it can be proven that g(t) defined by (2) posseses at most an algebraic singularity at the origin:

Theorem
If $(K_j)^2 \leq K^*$ and $\lambda_j / j^\beta \to 1$, $\beta > 1$, as $j \to \infty$, then $g(t) = O(t^{-(1/\beta)})$ as $t \to 0$.

When dealing with one-dimensional models of PN-diodes and provided C and μ are constant in the N-region and the P-region, respectively, the eigenvalues λ_j are the zeros of

$$\frac{\gamma_+}{\mu_+} \frac{\tan[\kappa_+(\lambda) l_N]}{\kappa_+(\lambda)} + \frac{\gamma_-}{\mu_-} \frac{\tan[\kappa_-(\lambda) l_P]}{\kappa_-(\lambda)} = 0 \qquad (11)$$

where the indices "+" / "-" denote the N-region / P-region, respectively, l_N and l_P are the scaled lenghts (one spatial dimension!) of these regions and

$$\kappa = \sqrt{\frac{|\lambda| - \frac{1}{\tau}}{\mu}} \qquad (12)$$

K_j can be computed from

$$K^2 \left[l_N \frac{1}{\gamma_+ \sin^2(\kappa_+ l_N)} - \frac{1}{\gamma_+ \kappa_+} \frac{\cos(\kappa_+ l_N)}{\sin(\kappa_+ l_N)} + l_P \frac{1}{\gamma_- \sin^2(\kappa_- l_P)} - \frac{1}{\gamma_- \kappa_-} \frac{\cos(\kappa_- l_P)}{\sin(\kappa_- l_P)} \right] = 2 \qquad (13)$$

It is easy to see, that $\lambda_j = O(j^2)$ as $j \to \infty$ and that $(K_j)^2$ stays bounded as $j \to \infty$. Hence, in the case of one-dimensional modelling of PN-diodes the physically observable constant-current phase has been justified from a purely mathematical point of view.

(*) The proofs of the cited theorems can be found in [UNTERREITER, 1991].

REFERENCES

[MARK,RING,SCH,1990]
P.Markowich, C.Ringhofer, C.Schmeiser : "Semiconductor Equations", Springer, Vienna.

[MILLER,UNTERREITER,1991]
R.K.Miller , A.Unterreiter : "On an integral equation arising in the description of the switching behaviour of PN-diodes", submitted for publication at the Journal of Integral Equations.

[vanROOSBROECK,1950]
W. vanRoosbroeck : "Theory of flow of electrons and holes in germanium and other semiconductors", Bell system Tech J, 29, 560.

[SCH,UNT,WEISS,1991]
C.Schmeiser,A.Unterreiter,R.Weiss. "The switching behaviour of PN-diodes in the case of low injection", in preparation.

[UNTERREITER,1991]
A.Unterreiter: "The switching behaviour of PN-diodes in the case of low injection", Thesis, TU Vienna.

A NOVEL COMPUTATIONAL PARADIGM: MUCH MORE EFFICIENT THAN VON NEUMANN PRINCIPLES

R.W. Hartenstein, H. Reinig, M. Riedmüller, K. Schmidt

Universitaet Kaiserslautern, Fachbereich Informatik, Bau 12
Postfach 3049, D-W-6750 Kaiserslautern, Germany
phone: (+49-631) 205-2606, Fax: (+49-631) 205-3[illegible]00

Abstract

Computers (based on von Neumann principles) are extremely inefficient. That's why this paper introduces a novel computational paradigm based on new hardware machine principles. Such machines, called "xputers" avoid most of the bottlenecks known from (von Neumann) computers, so that a hardware efficiency is obtained which is higher by several orders of magnitude. By means of a few algorithm examples the new paradigm will be introduced as a new programming paradigm, which is data-procedural (which is more direct than the control-procedural von Neumann paradigm). Finally the paper gives a survey on the novel application development environments needed for xputers and their advantages over such tools for computers. Such application support for xputers includes two alternative source levels: high level programs, or very high level algorithm specifications.

The Machine Paradigm of Xputers

In programming and coding von Neumann machines their basic paradigm causes massive overhead (control flow overhead, addressing overhead and other kinds of overhead [1]), which eats up most of the processor's throughput and its primary memory bandwidth. Due to the control-driven operation principles by multiplexer-based instruction sequencing the individual processor permits only very limited intra-ALU parallelism [1]. Parallel computer systems suffer from high hardware cost and from massive inter-processor communication overhead [2]. In addition to that most of the processors are idling (a very few applications are exceptions from that). Data flow machines suffer from massive token flow overhead needed for arbitration and from other kinds of overhead [2] and debugging is extremely difficult.

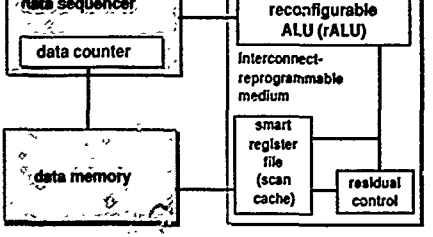


fig. 1: basic blocks of xputers

Xputers use a data sequencer (instead of an instruction sequencer) including a powerful address generator to avoid addressing overhead and control flow overhead whenever possible (see also fig. 1). Instead of an ALU driven by an instruction sequencer, xputers use a reconfigurable ALU (called rALU), such that powerful compound operators exhibiting intra-ALU parallelism may be defined at compile time. More hardware details have been published elsewhere [1-3]. A smart register file (called *scan cache*, being a size adjustable window to memory space) including an intelligent data memory interface supports new compile time optimization strategies being more efficient than those which can be mapped on von Neumann hardware only. Algorithm execution examples within other sections of this paper will give a flavour of the high efficiency of this novel machine organization.

For computers the RAM (random-access memory) is the *central technology platform* which provides flexibility and universality. For xputers, however, the flexibility and universality is derived from using *interconnect-reprogrammable media* (compare fig. 1), also called *field-programmable media*, which are commercially available from a billion US-dollar niche of the integrated circuit market (1990: world-wide).

Programming and Compilation for Xputers

The Xputer machine paradigm is also visible from a programmer's point of view. Figures 2 and 3 show two simple algorithm examples, which will be used to illustrate programming for xputers, execution of programs on xputers, as well as the compilation for xputers. Fig. 2a shows the textual notation of a simple recurrent 8 step loop. The right side of fig. 2b shows its equivalent graphic notation: a signal flow graph (SFG) which reveals the regular data dependencies. For programming an xputer from such a specification the following code elements are needed (to be generated by a compiler): a *rALU subnet* (*compound operators*: saving memory cycles, since intermediate results are not stored) has to be derived (e.g. see fig. 2b: by just picking an iteration from the SFG); ad data map has to be derived (2-dim. data map example in fig. 2c: mapping the SFG onto a grid); the scan cache size has to be selected (e.g. 1-by-4 words: see fig. 2d); and an address sequence (called *scan pattern*: see fig. 2d) has to be selected.

Fig. 2d also illustrates xputer operation in executing the algorithm example from figures 2a and 2b. First the data sequencer makes the scan cache jump onto the leftmost column of the data map. Fig. 2d also illustrates the auto-apply mode and the auto-copy mode. This means that, whenever the scan cache is placed onto a particular memory location, it automatically (i.e. without needing a controller, thus avoiding control overhead) invokes operation of the rALU subnet currently selected, as well as a cache/memory communication cycle. Since a register has been added to the rALU subnet to save c[i], only four memory read cycles are needed (see left side of fig. 2d). Due to the scan pattern in our example this is repeated 8 times until finally the scan cache arrives at the rightmost position of the data map. The tagged control word (TCW) found there tells where to store the final result and provides the linkage code to select the next scan pattern etc. So o[illegible] out of 34 memory semi cycles has been used for control, what we call *sparse control* or *residual control* (also compare fig. 1).

Fig. 3 shows another simple algorithm example: a 2-dimensional filtering example where a simple video scan is used as a scan pattern (fig. 3b) to scan a 2-dim. data map (fig. 3b) by a 3-by-3 word scan cache (fig. 3a) Fig 3c shows the very powerful compound operator, which needs relatively few hardware, since smaller data path width (8 bits) could be traded for more operators This illustrates the high flexibility of xputers By code analysis it has been found, that in the von Neumann version of the same algorithm

(on Vax-11/750) 93% of computation time is used for address computation. The MoM [2] xputer version avoids this addressing overhead by its hardwired address generator [3] running concurrently (MoM stands for Map-oriented Machine), where the 2-dim. memory organization permits even better performance. A spreadsheet representation of the data map [4] supports a highly intuitive user interface for xputer applications.

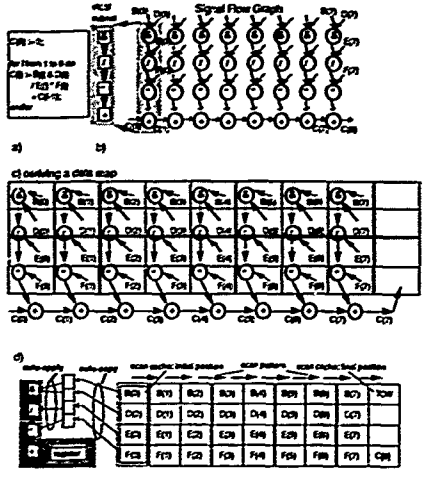


fig. 2: A simple algorithm example: a) textual, b) graphic specification, c) deriving a data map, d) xputer operation illustration

High Performance Applications

Fig. 4 shows a number of acceleration factors having been obtained experimentally from a number of algorithm implementations on the MoM-1 and MoM-2 xputers, compared to implementations on a Vax-11/750. These results show, that for a number of application areas (Digital Signal Processing, image processing, computer graphics, uniform equation systems, etc.) xputers are competitive to many ASIC solutions, to implementations on special processors (e.g digital signal processors), or to implementations on supercomputers. A so called *retargeting* feature, being available commercially, permits a direct path from an implementation on a programmable xputer to a very cheap customized silicon solution: machine code can be directly submitted for gate array fabrication, such that no expensive ASIC design process is needed [2].

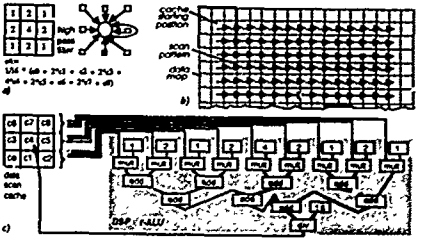


fig. 3: A 2-dimensional filtering example

Conclusions

A novel high performance machine paradigm of xputers has been introduced, which for a very large class of algorithms is by several orders of magnitude more efficient than the von Neumann paradigm. The data-procedural principles of operation, which are more intuitive and more direct than traditional principles, have been illustrated by 2 examples. two experimental xputer architecture examples (the MoM-1 and the MoM-2) have been implemented by breadboard techniques. Highly promising acceleration factors for a number of algorithms have been obtained experimentally by comparing MoM implementations versus implementations on Vax-11/750. The results show, that xputers are competitive to many ASIC solutions as well as to many algorithm implementations on supercomputers. A third xputer architecture (the MoM-3) stressing flexibility and universality is currently being implemented at Kaiserslautern.

application example	acceleration factor
CMOS design rule check	>2000
sup. electrical rules check	>300
Lee routing without obstacles	>160
with obstacles	>800
2-dimensional filtering (3 by 3)	>300
vector-matrix multiplication	150

fig. 4: acceleration factors obtained experimentally (xputer implementation versus optimized Vax-11/750 implementation

References

[1] R. Hartenstein et al: A Novel Paradigm of Parallel Computation and its Use to Implement Simple High Performance Hardware; Intl. Conference on Information Technology, Japan, Oct. 1990

[2] R. Hartenstein et al: Xputers: An Open Family of Non-von Neumann Architectures; internal report; University of Kaiserslautern, 1989

[3] A. Hirschbiel: A Novel Processor Architecture Based on Auto Data Sequencing and Low Level Parallelism; Ph. D. Thesis; University of Kaiserslautern, March 1991

[4] M. Weber: An Application Development Method for Xputers; Ph. D. Thesis, University of Kaiserslautern, Dec. 1990

NEW METHOD OF DYNAMIC LINEAR SYSTEMS ANALYSIS BASED ON THE MEMORYLESS ALGORITHMS OF CONVOLUTION CALCULATIONS

Janusz Zarębski
Institute of Radioelectronics
Merchant Marine Academy of Gdynia, Poland

Abstract – In the paper the memoryless method of analysis of dynamic linear systems described by convolution integral is presented. The memoryless method is applicable if one of the convolved functions is a sum of exponential terms with real or complex coefficients. The algorithm based on the proposed method is numerically simpler then the known, direct algorithms of convolution in the time domain.

1. INTRODUCTION

The natural way for the derivation of the output signal $y(t)$ of any linear, time-invariant system is to calculate numerically the convolution integral

$$y(t) = k(t) * x(t) = \int_0^t k(t-\tau) \cdot x(\tau) d\tau. \quad (1)$$

where: $k(t)$-the pulse response of the analysed system, $x(t)$-the input signal, for the sequence of the time sample values $t_n (n=1,2,\ldots,N)$. The direct algorithms in the time domain for convolutions are numerically complex, because the number of multiplications in the realization of these algorithms for n samples is proportional to n^2. Usually, in applications numerous techniques based on the Fourier or Laplace transformations are used [1] If one of the convolved functions, e.g. $k(t)$ is of the form

$$k(t) = \sum_{i=1}^{L} A_i \cdot \exp(-\alpha_i \cdot t), \quad (2)$$

then the memoryless algorithm described in the paper [2] can be used In the paper the memoryless algorithm in the time domain for the analysis of a larger class of linear systems is presented

2. THE BASIS OF THE MEMORYLESS ALGORITHMS

If in eq.(2) coefficients α_i, for $i=1,2,\ldots,L$, denote the real numbers then to solve eq.(1) in the n-th step of calculations the following memoryless algorithm can be used [3]

$$y_n^i = \zeta^i \cdot y_{n-1}^i + h_1^i \cdot x_{n-1} + h_2^i \cdot x_n . \quad (3)$$

$$y_n = \sum_{i=1}^{L} y_n^i. \quad (4)$$

where $\zeta^i = \exp(-\alpha_i \cdot \Delta)$, (5)

$$h_1^i = \frac{\Delta}{2} \cdot k_1^i . \quad (6)$$

$$h_2^i = \frac{\Delta}{2} \cdot k_0^i . \quad (7)$$

The memoryless algorithm can also be formulated for the case in which the $k(t)$ function is given of the form of the sum of trigonometric functions. Idea of such algorithm will be explain for the simple case, when

$$k(t) = B \cdot e^{-a \cdot t} \cdot \cos\omega t. \quad (8)$$

Because of

$$Be^{-at}\cos\omega t = \frac{B}{2} \cdot \exp[-(a-j\omega)t] + \frac{B}{2} \cdot \exp[-(a+j\omega)t], \quad (9)$$

than, after taking advantage of linear properties of the convolution integral, we receive

$$y(t) = x(t) * \frac{B}{2} \exp[-(a-j\omega)t] + x(t) * \frac{B}{2} \exp[-(a+j\omega)t] = y_1(t) + y_2(t), \quad (10)$$

where

$$y_1(t) = y_2^*(t) \quad (11)$$

that means

$$\mathrm{Re}\,[y_1(t)] = \mathrm{Re}\,[y_2(t)] \quad (12)$$

$$\mathrm{Im}\,[y_1(t)] = -\,\mathrm{Im}\,[y_2(t)] \quad (13)$$

So, in the n-th step of calculations

$$y_n = 2\,\mathrm{Re}(y_{n1}) = 2\,\mathrm{Re}(y_{n2}). \qquad (14)$$

For the clarity of algorithm, in the further consideration, let's take

$$y_n = 2\,\mathrm{Re}(y_{n1}). \qquad (15)$$

Thus, the parameters in eq(2) can be described by the following dependences

$$A = \frac{B}{2}, \qquad (16)$$

$$\alpha = C\left(\frac{a}{a^2+\omega^2} + j\,\frac{\omega}{a^2+\omega^2}\right)^{-1}, \qquad (17)$$

and from eq (5) we have

$$\zeta = e^{-a\Delta}(\cos\omega\Delta + j\sin\omega\Delta). \qquad (18)$$

After including eqs.(3,4,6,7,14-18), for the n-th step of calculations, the memoryless algorithm is of the form

$$\begin{cases} \mathrm{Re}(y_n^1) = q\cdot[\mathrm{Re}(y_{n-1}^1)\cdot\delta_1 - \mathrm{Im}(y_{n-1}^1)\cdot\delta_2 + \\ \quad + \mathrm{Re}(h_1)\cdot x_{n-1} + \mathrm{Re}(h_2)\cdot x_n \\ \mathrm{Im}(y_n^1) = q\cdot[\mathrm{Im}(y_{n-1}^1)\cdot\delta_1 + \mathrm{Re}(y_{n-1}^1)\cdot\delta_2 + \\ \quad + \mathrm{Im}(h_1)\cdot x_{n-1} + \mathrm{Im}(h_2)\cdot x_n \\ y_n = 2\,\mathrm{Re}(y_n^1). \end{cases} \qquad (19)$$

where

$$h_1 = q\cdot(B\cdot\frac{\Delta}{4}\cdot\delta_1 + j\,B\cdot\frac{\Delta}{4}\cdot\delta_2), \qquad (20)$$

$$h_2 = B\cdot\frac{\Delta}{4}, \qquad (21)$$

and $q = e^{-a\cdot\Delta}$, $\delta_1 = \cos\omega t$, $\delta_2 = \sin\omega t$.

In the same way we can formulate the memoryless algorithm for the function k(t) of the form of $\sin\omega t$.

3. THE MAIN ALGORITHM

The memoryless algorithm from section 2 can be generalized for the Fourier representation of function of k(t) of the form

$$k(t) = \sum_{i=1}^{M_1} A_i \exp(-a_i\cdot t)\cos\omega_i t + \sum_{k=1}^{M_2} B_k \exp(-b_k t)\sin\omega_k t. \qquad (22)$$

In this case the consecutive samples of the convolution y(t) are derived according to equations

$$\begin{cases} y_n = 2\cdot\left[\sum_{i=1}^{M_1}\mathrm{Re}(y_{cn}^i) + \sum_{k=1}^{M_2}\mathrm{Re}(y_{sn}^k)\right] \\ \mathrm{Re}(y_{cn}^i) = q_1^i\cdot[\mathrm{Re}(y_{c(n-1)}^i)\cdot\delta_1^i - \mathrm{Im}(y_{c(n-1)}^i)\cdot\delta_2^i] \\ \quad + \frac{A_i\cdot x_{n-1}}{4}\cdot\Delta\cdot q_1^i\cdot\delta_1^i + \frac{A_i\cdot x_n}{4}\cdot\Delta, \\ \mathrm{Im}(y_{cn}^i) = q_1^i\cdot[\mathrm{Im}(y_{c(n-1)}^i)\cdot\delta_1^i + \mathrm{Re}(y_{c(n-1)}^i)\cdot\delta_2^i] \\ \quad + \frac{A_i\cdot x_{n-1}}{4}\cdot\Delta\cdot q_1\cdot\delta_2, \\ \mathrm{Re}(y_{sn}^k) = q_2^k\cdot[\mathrm{Re}(y_{s(n-1)}^k)\cdot\delta_2^k - \mathrm{Im}(y_{s(n-1)}^k)\cdot\delta_2^k] \\ \quad + \frac{B_k\cdot x_{n-1}}{4}\cdot\Delta\cdot q_2\cdot\delta_2^k, \\ \mathrm{Im}(y_{sn}^k) = q_2^k\cdot[\mathrm{Im}(y_{s(n-1)}^i)\cdot\delta_2^k + \mathrm{Re}(y_{s(n-1)}^k)\cdot\delta_2^k] \\ \quad + \frac{B_k\cdot x_{n-1}}{4}\cdot q_2^k\cdot\Delta\cdot\delta_2^k + \frac{B_k\cdot x_n}{4}\cdot\Delta. \end{cases} \qquad (23)$$

where $\delta_1^i = e^{-a_i\cdot\Delta}$, $q_2^k = e^{-b_k\cdot\Delta}$.

Series of numerical calculations performed with the use of computer program based on the described algorithm confirm the correctness and effectiveness of the memoryless method.

4. SUMMARY

In the paper the new algorithm of the convolution integral calculation is proposed. This algorithm is applicable for the special form of the convolved functions. The principal advantages of the proposals algorithms consists in theirs numerical simplicity and in theirs "memoryless" properties, that means, for the calculation of any output sample y_n only the actual input sample value x_n and the preceding output sample value y_{n-1} are sufficient.

REFERENCES

1 Agarval R.C., Cooley J.W.: New Algorithm for Digital Convolutions, IEEE Trans on Acoust., vol 1977 N.4., p.392.

2 Janke W., Zarębski J · Fast Algorithm for the Special Case Convolution and Deconvolution Calculations, IEEE Intern. Symp. on Circuits and Systems, New Orleans, Louisiana, 1990, vol. 3, p. 2377.

A REGULARIZED SOLUTION FOR DECONVOLUTION OF SIGNALS WITH EDGES BY AN ADAPTIVE ITERATIVE VERSION OF TIKHONOV-MILLER METHOD

CARMEN SANCHEZ AVILA
Department of Applied Mathematics
E.T.S.I. Telecomunicación. U.P.M.
Ciudad Universitaria s/n
28040 Madrid. SPAIN.

Abstract: An adaptive regularized iterative algorithm which solves the "ill-posed" problem of deconvolution of signals with edges in a numerically stable way by incorporating an Adaptive Projection Operator is described. Considering the joint detection-estimation character that edges deconvolution problem have, we introduce an adaptively contracted (projection) selection operator to detect the sharp changes in intensity (2-D) or the edges (1-D) of the solution, which can be combined with iterative algorithms based in the Tikhonov-Miller method. A numerical example shows the improvement of this new adaptive method proposed.

I. INTRODUCTION: DESCRIPTION PROBLEM

In signal deconvolution the ultimate goal is the recovery of the original signal from a degraded version

$$\int_a^b k(x,t)f(t)dt = g(x) \qquad a \le x \le b \qquad (1)$$

or formally:

$$H f = g \qquad (2)$$

We will consider the discrete version of (2)

$$H f = g \qquad (3)$$

In addition, the degraded signal is nearly always corrupted by random noise. We model our noisy degraded signals as follows:

$$H f + n = g \qquad (4)$$

where H is an impulse response. We will here concetrate on situations in which H is assumed known, and the abrupt character of f is also known a priori.

Strictly speaking, (4) establishes a detection plus estimation problem. First, we have to identify the edges of f (original signal), then, to estimate the values of this sharp changes.

Is well-known that this problem corresponds to resolve a Fredholm integral equation of first kind, and, in Hilbert spaces is generally an "ill-posed" problem. We will say that a problem is "well-posed", in the Hadamard sense [1], if verifies the following conditions: i) there is a solution for each g, ii) the solution is unique, and iii) the solution is a continous function of the data g, that is the problem is stable to small fluctuations of g. If some of these characteristics is not satisfied, we will say that the inverse problem is "ill-posed". Then, in our case a) a solution in the classic sense doesn't exist, and it's necessary to define what we'll call an approach solution to the problem (2), b) the problem is not stable as far as g_1 and g_2 are close elements and however their respective solutions f_1 and f_2, may be substantially different.

In recent years iterative signal restoration, an iterative signal deconvolution in particular, has been given considerable attention.

However, most of the existing iterative algorithms are derived without explicitly taking into account the presence of noise in the degraded signal. As a result excessive noise amplification may occur and heavy ringing effects will be visible when the number of iterations increases. It can be shown that both effects are due to the "ill-posedness" or "ill-conditionedness" of the deconvolution problem [1]. In order to solve the "ill-posed" signal deconvolution problem, a priori knowledge about the original signal has to be included in the derivation of the deconvolution algorithm. Such an approach is commonly referred to as "regularization" [1].

Our approach is to find a regularized solution to an "ill-conditioned" system of equations. The regularization approach we take here is based in the Tikhonov-Miller method, which we will revise in section II. In section III we propose a way to do this method adaptive. In section IV we present a numerical example which shows the advantages of this new method.

II. THE TIKHONOV-MILLER METHOD

The theory of "ill-posed" problems and the methods to solve them began to be extensively developed after fundamental works by Tikhonov [2-3]. The most important discovery in this field was the concept of approximate solutions of "ill-posed" problems [2], the means of finding such solutions being based on the concept of regularizing algorithms [3]. Tikhonov has suggested the following scheme for costructing a regularizing algorithm.

Let us consider the "ill-posed" problem (2), where f and $g \in H$ (H is a Hilbert space finite-dimensional), and H is a continuous linear one-to-one operator from H to H
Let us consider the functional:

$$M_\alpha(f) = \| H f - g \|_2^2 + \alpha \| f \|_2^2$$

where α is called: regularization parameter. It can be seen that under the above conditions the functional $M_\alpha(f)$ (for any $\alpha>0$) is very convex and it will have a lower bound on H only at the point f_α. This extremal f_α of functional $M_\alpha(f)$ is found from the solution of

$$H^T H f + \alpha f = H^T g$$

Additionally from the regularization theory [1], it is well known that physically meaningful solutions to "ill-posed" problems can be obtained by incorporating a priori information about the original data or the noise into the solution method.

In signal deconvolution we usually have available an estimate of the norm of the noise present in the degraded signal; so

$$\| g - H f_e \|_2^2 \le \varepsilon \qquad (5)$$

The bound ε is related to the amount of the noise present, and it is assumed a priori known.

The Tikhonov-Miller regularization [4], provides a second kind of a priori knowledge. To regulate locally the tradeoff between the noise magnification error and the regularization error, we propose to impose an upper bound on the norm of the filtered signal $L f_e$, where L is the linear regularization operator (here, L is a high-pass filter)

$$\| L f_e \|_2^2 \le E \qquad (6)$$

The bound E is assumed to be (approximately) known a priori.

Following the Miller regularization approach [4] we combine (5) and (6) into a single quadrature formula

$$\Phi(f_e) = \| g - H f_e \|_2^2 + \alpha \| L f_e \|_2^2 \qquad (7)$$

The regularization parameter has the fixed value $\alpha=(\varepsilon/E)^2$. Among the solutions satisfying (7), a physically reasonable choice is the solution f_e^m, called the Miller regularized solution, which minimizes the functional $\Phi(f_e)$. The solution of this minimization problem is given by the normal equations

$$(H^T H + \alpha L^T L) f_e^m = H^T g \qquad (8)$$

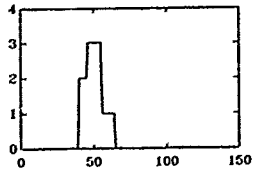


Fig. 1. *The original signal.*

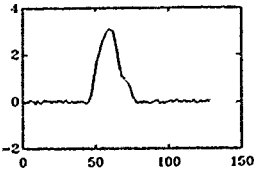


Fig. 2. *The degraded signal. SNR = 23 dB.*

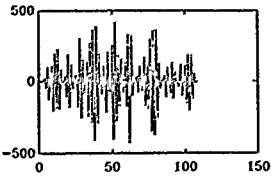


Fig. 3. *Moore-Penrose pseudoinverse*

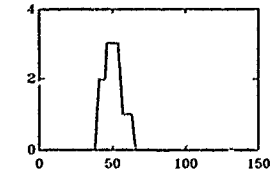


Fig. 4. *The final result with the proposed adaptive Tikhonov-Miller method. It = 22.*

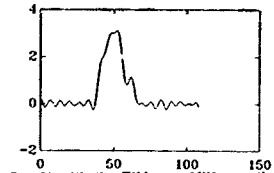


Fig. 5. *Result with the Tikhonov-Miller method. It=120.*

Here, the solution f_e^m is approximated by using an iterative method, which simultaneously offers the possibility of imposing a deterministic constraint on the solution.

We rewrite (8) as:

$$f_e^{m+1} = (I - \alpha\beta L^T L) f_e^m + \beta H^T (g - H f_e^m) = C(f_e^m) \quad (9)$$

where β is an useful relaxation parameter to insure the contractiveness of the mapping C.

III. AN ADAPTIVE TIKHONOV-MILLER METHOD

The knowledge about the abrupt character of the original signal would be equivalent to introducing a square selection matrix P in (4).

$$H P f + n = g \quad (10)$$

We can considered P as a Projection Operator, such f is a fixed point, i.e. $P f = f$. We define the Projection Operator P as

$$P = B^{-1} P_D B$$

where B is a basis change matrix, such that Bf represents f in the basis

$$B = \{ \delta(n-i)-\delta(n-1-i) \} \qquad \forall i$$

P_D will be a diagonal matrix, with all the elements equal to zero with the exception of those corresponding to nonzero values of $B f$. Unfortunately, P_D is not available in practice. However we can try to define adaptively in each iteration. So:

$$P^m = B^{-1} P_D^m B$$

For the new system (10), the algorithm (9) is given by

$$f_e^{m+1} = (I - \alpha\beta (LP)^T (LP)) f_e^m + \beta P^T H^T (g - H P f_e^m)$$

Considering that $P^T = P$ and that we need to define a selection matrix at each step, we arrive at:

$$f_e^{m+1} = (I - \alpha\beta P^m L^T L P^m) f_e^m + \beta P^m H^T (g - H P^m f_e^m)$$

We will assume here a contraction of the projection domain (i.e., the nonzero elements of P_D^{m+1} must be a subset of those of P_D^m). Assuming such a contraction, we can write:

$$f_e^{m+1} = P^m (f_e^m - \alpha\beta L^T L P^m f_e^m + \beta H^T (g - H P^m f_e^m))$$

We will define P_D^m after obtaining f_e^m by means a threshold procedure proposed by Papoulis and Chamzas in an extrapolation context [5]

$$P_D^m(i,i) = \begin{cases} 1 & \text{if } | f_e(i) - f_e(i-1) | > T_m \\ 0 & \text{if } | f_e^m(i) - f_e^m(i-1) | \le T_m \end{cases}$$

where:

$$T_m = \max \{ T_{m-1} , \gamma \min | f_e^k - f_e^{k-1} | \; k \in A \}$$

γ being a empirical constant, $0 < \gamma < 1$, and A indicating the set of points selected by T_{m-1}.

IV. A NUMERICAL EXAMPLE

We present the results of a computer simulation to illustrate the performance of the method $H(m,n)$ corresponds to the values of the impulse response $h(m-n)$ of a low-pass filter. Fig.1 shows the original signal f, where the only nonzero differences are $f(40)-f(39)=2$, $f(46)-f(45)=1$, $f(56)-f(55)=-2$, and $f(65)-f(64)=-1$. Fig 2 shows the degraded signal $g = H f + n$, where n is a zero mean Gaussian white noise with $\sigma=0.02$. Fig 3 demostrates how the applications of a classic method (non regularized) the Moore-Penrose pseudoinverse does not provide a useful result. Fig 4 shows the final result. f_e^{22}, applying the proposed method with $f_e^0 = 0$, and $\beta = 1.5$. Finally, Fig 5 shows the result obtained applying the version iterative of Tikhonov-Miller method (9) without the Adaptive Operator. We can see the improvement introduced by the new adaptive regularized method.

REFERENCES

[1] Tikhonov, A.N. and Arsenin V.Y., *Solutions of Ill-Posed Problems*, Wiley, Washington, 1977.

[2] Tikhonov, A.N., *Dokl. Akad. Nauk. SSSR* 151, 501,(1963)

[3] Tikhonov, A.N., *Dokl. Akad. Nauk. SSSR* 153, 49, (1963)

[4] Miller, K., *SIAM J. Math Anal.*, vol 1, 1970, pp 52-74

[5] Papoulis, A. and Chamzas, C., *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol ASSP 27, no 5, 1979,pp 492-500

FEASIBILITY OF LYAPUNOV FUNCTIONS FOR POWER SYSTEM TRANSIENT STABILITY ANALYSIS BY THE CONTROLLING U.E P. METHOD

A. M. ESKİCİOĞLU
Computer Engineering Department
Middle East Technical University
06531 Ankara, Turkey

O. SEVAİOĞLU
Electrical & Electronics Eng. Dept.
Middle East Technical University
06531 Ankara, Turkey

Abstract - Recently a great deal of improvement has been achieved in reducing the inherent conservativeness of the direct method of Lyapunov in its application to the transient stability analysis of multi-machine power systems. One major discovery is the recognition of the fault location as a critical factor for the determination of the boundary of the stability region. The controlling u.e p. method takes into consideration the relevant unstable equilibrium points to construct Lyapunov surfaces, local approximations to the stability boundary, giving more accurate estimates In this paper, three common Lyapunov functions are used in a comparative evaluation to determine their feasibility for the local Lyapunov surface approach. It is shown that the usefulness of different Lyapunov functions varies greatly in various cases of instability occurring in the applications.

I. INTRODUCTION

There has been a substantial amount of research in the use of direct methods for the transient stability analysis of multi-machine power systems in the last two decades [1]-[4]. Among the various techniques proposed, Lyapunov's direct method has attracted much attention and proved to be a promising tool of analysis for off-line and on-line studies. Initially, the results obtained using Lyapunov's direct method were, in general, practicably too conservative in spite of the fact that numerous Lyapunov functions were developed and tried The main reason for this conservativeness was later understood to be determination of the stable region in the state space independent of the fault location. In this original approach, commonly called the closest u e.p method [4], a global approximation to the separatrix, the stability boundary, is obtained by constructing a single Lyapunov surface using the unstable equilibrium point which gives the minimum value of the chosen Lyapunov function.

To alleviate the inherent conservative nature of the closest u e.p. method, recent studies were focused on identifying the unstable equilibrium point corresponding the fault under consideration, and employing it for the construction of the local Lyapunov surface. Referred to as the controlling u.e.p. method [4], the new approach provides a local approximation to the separatrix, producing much improved estimates for critical clearing times. The fundamental problem with this method is the difficulty in determining the controlling u e p , and several methodologies have been proposed for a solution [5]-[10] It is evident that both the closest and the controlling u e p methods give the same result only when the closest unstable equilibrium point happens to be the relevant unstable equilibrium point.

The controlling u e p. method exhibits a close similarity to the method of tangent hyperplanes [11] or the method of tangent hypersurfaces [12]. All make use of the so-called type 1 points in constructing local approximations.

The use of local Lyapunov surfaces for the determination of transient stability regions is reported in a previous work [13]. A commonly used Lyapunov function was taken and the results obtained by the local and global approximations were compared Similar investigations with different Lyapunov functions are also available in the literature The major aim of this study is to make a comparative evaluation of three well-known Lyapunov functions with the purpose of determining their feasibility for transient stability analysis using the controlling u e.p method.

II. POWER SYSTEM MODEL

In the following analysis, the classical power system model is used In this model, the state equations of motion are mathematically expressed as

$$\begin{aligned}\dot{\delta}_i &= \omega_i, \qquad i = 1,\ldots,n\\ \dot{\omega}_i &= 1/M_i[P_{mi} - D_i\omega_i - E_i^2 G_{ii}\\ &\quad - \sum_{\substack{j=1\\ j\neq i}}^{n} E_i E_j Y_{ij}\cos(\delta_i - \delta_j - \Theta_{ij})],\ \ i = 1,\ldots,n\end{aligned}$$

where

δ_i	=	rotor angle of the ith machine
ω_i	=	rotor speed of the ith machine
τ	=	$2\pi f t$ (f = system frequency, t = time in seconds)
H_i	=	inertia constant of the ith machine
M_i	=	$(4\pi f)\, H_i$
D_i	=	$(2\pi f)\, d_i$ (d_i = damping coefficient of the ith machine)
P_{mi}	=	mechanical power input to the ith machine
E_i	=	constant voltage behind the direct-axis transient reactance
Y_{ij}	=	modulus of the ijth element of the reduced system admittance matrix
Θ_{ij}	=	argument of the ijth element of the reduced system admittance matrix
$(\dot{})$	=	$d/d\tau$

It is stated that within the Lyapunov context, the correct mathematical model should be described by the relative, and not by the absolute, rotor angles and speeds [1] Taking the nth machine as reference and defining the state vector as

$$x = [\delta_{1n}, \delta_{2n}, \quad, \delta_{n-1,n}, \omega_{1n}, \omega_{2n}, \ldots, \omega_{n-1,n}]^T$$

the above equations lead to

$$\begin{aligned}\dot{\delta}_{in} &= \omega_{in}, \quad i = 1, \quad, n-1\\ \dot{\omega}_{in} &= f_i(\delta, \omega)\\ &= 1/M_i(P_{mi} - E_i^2 G_{ii}) - 1/M_n(P_{mn} - E_n^2 G_{nn})\\ &\quad -1/M_i[E_i E_n Y_{in}\cos(\delta_{in} - \Theta_{in})\\ &\quad + \sum_{\substack{j=1\\ j\neq i}}^{n-1} E_i E_j Y_{ij}\cos(\delta_{in} - \delta_{jn} - \Theta_{ij})]\\ &\quad +1/M_n[\sum_{j=1}^{n-1} E_j E_n Y_{jn}\cos(\delta_{jn} + \Theta_{jn})]\\ &\quad -C\omega_{in}, \quad i = 1, \quad, n-1\end{aligned}$$

where $C = D_i/M_i$ for all i

III. THREE COMMON LYAPUNOV FUNCTIONS

The three Lyapunov functions considered for comparison are widely used and frequently discussed in the literature. They are

$$V_1 = V_{11} + V_{12} + V_{13} \qquad [14],$$

where

$$V_{11} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 1/2M_iM_j(\omega_{in}-\omega_{jn})^2$$

$$V_{12} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{ij}(\delta_{in}-\delta_{jn}+\delta^s_{jn}-\delta^s_{in})$$

$$V_{13} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_iE_jB_{ij}M\left[\cos(\delta_{in}-\delta_{jn})-\cos(\delta^s_{in}-\delta^s_{jn})\right]$$

$$V_2 = V_{21} + V_{22} + V_{23} \qquad [15],$$

where

$$V_{21} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} M_iM_j(\omega_{in}-\omega_{jn})^2$$

$$V_{22} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} ME_iE_jB_{ij}(\delta_{in}-\delta_{jn}-\delta^s_{in}+\delta^s_{jn})\sin(\delta^s_{in}-\delta^s_{jn})$$

$$V_{23} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} ME_iE_jB_{ij}\left[\cos(\delta_{in}-\delta_{jn})-\cos(\delta^s_{in}-\delta^s_{jn})\right]$$

$$V_3 = V_{31} + V_{32} + V_{33} \qquad [16],$$

where

$$V_{31} = 1/2\sum_{i=1}^{n} M_i\omega_{in}^2$$

$$V_{32} = -\sum_{i=1}^{n}(P_{mi}-E_i^2G_{ii})(\delta_{in}-\delta^s_{in})$$

$$V_{33} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_iE_jB_{ij}\left[\cos(\delta_{in}-\delta_{jn})-\cos(\delta^s_{in}-\delta^s_{jn})\right]$$

In the above expressions, superscript s denotes value of the angle at the stable equilibrium point, and

$$C_{ij} = M_j(P_{mi}-E_i^2G_{ii}) - M_i(P_{mj}-E_j^2G_{jj})$$

$$M = \sum_{i=1}^{n} M_i$$

B_{ij} = susceptances

It should be noted that V_1 has already been used in [13]. However, in the modulus of the ij-th element of the reduced system admittance matrix transfer conductances were not assumed to be zero, producing rather different results especially for cases of multi-generator instability. It should also be noted that the Lyapunov function given in [5], which is also common, has been shown to be identical to V_1 and hence can be assumed to give the same results reported for V_1.

To account for transfer conductances, G_{ij}, the expressions for the first Lyapunov function can be 'corrected' by assuming a linear faulted trajectory in the state space and, hence, adding a fourth term

For V_1 [17],

$$V_{14} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\{M_j\sum_{k=1}^{n} E_iE_kG_{ik}[\sin(\delta_{in}-\delta_{kn})-\sin(\delta^s_{in}-\delta^s_{kn})]/$$
$$[\delta_{in}-\delta_{kn}-(\delta^s_{in}-\delta^s_{kn})]$$
$$-M_i\sum_{k=1}^{n} E_jE_kG_{jk}[\sin(\delta_{jn}-\delta_{kn})-\sin(\delta^s_{jn}-\delta^s_{kn})]$$
$$[\delta_{in}-\delta_{jn}-(\delta^s_{in}-\delta^s_{jn})]/[\delta_{jn}-\delta_{kn}-(\delta^s_{jn}-\delta^s_{kn})]\}$$

A correction term for the Lyapunov function V_2 is derived in [18] as

$$V_{24} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(M-2M_j)E_iE_jG_{ij}$$
$$\{\cos(\delta^s_{in}-\delta^s_{jn})(\delta^s_{in}-\delta^s_{jn}-(\delta_{in}-\delta_{jn}))$$
$$-(\sin(\delta^s_{in}-\delta^s_{jn})-\sin(\delta_{in}-\delta_{jn}))\}$$

This term is obtained by the exact integration of a part of the non-integrable terms hither to neglected.

Finally, the fourth term for the Lyapunov function V_3 is taken to be [16]

$$V_{34} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_iE_jG_{ij}\left[\sin(\delta_{in}-\delta_{jn})-\sin(\delta^s_{in}-\delta^s_{jn})\right]$$

The effect of adding the above terms in the expressions for the given functions is also checked for the test systems used in this work to establish their validity.

IV. APPLICATION TO TEST SYSTEMS

Two power systems are used in the application of the controlling u e p method together with the closest u e.p method. The smaller is a 4-generator system [16] which has extensively been tested in numerous studies. The other system [19] is provided by the South of Scotland Electricity Board and has 12 generators.

Stability characteristics of the above systems are investigated by applying a 3-phase-to ground fault, one of the severest that a power system may be subjected to, at each of the generator terminal buses. The fault is assumed to be cleared without line tripping. Because of the properties of the classical model, only the first-swing instability is taken into consideration. The value for uniform damping is assumed to be zero. This is the usual choice as accounting for uniform damping improves the quality of systems stability but not its quantity [1]. However, the analysis presented in this paper was also carried out using a nonzero value for damping. Because of space limits, the results are not included here. As a passing remark, it can be said that they are slightly worse than those for zero damping.

4.1 Results for 4-Generator System

Three type 1 (unstable) equilibrium points are found for the system. They represent the modes of instability for generator 1, 2 and 3. The estimates computed for critical clearing times (CCT's) are listed in Table 1. CCT_1's and CCT_2's denote the CCT's estimated by the controlling u e p method and the closest u.e.p. method, respectively.

Table 1. Critical Clearing Times for 4-Generator System

(a) Estimates of the Critical Clearing Times by using V_1

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	137785[a]	0.557	0.500	-10 2	0.500	-10 2
2	170267	0.441	0.434	-1.9	0.391	-11.3
3	216719	0.491	0.483	-1.6	0.396	-19.3

(b) Estimates of the Critical Clearing Times by using V_2

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	137246[a]	0.557	0.511	-8.3	0.511	-8.3
2	160721	0 441	0.429	-2.7	0 402	-8.8
3	195373	0 491	0.476	-3.1	0.414	-15.7

(c) Estimates of the Critical Clearing Times by using V_3

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	1.6585[a]	0.557	0 498	-10.5	0.498	-10 5
2	2.1518	0.441	0.433	-1.8	0 390	-11.6
3	2.7239	0.491	0.483	-1 6	0 397	-19.3

[a]Minimum value of the Lyapunov function.

Examining the results in Table 1 leads to the primary conclusion that local Lyapunov surfaces yield always better approximations through which much improved CCT estimates are obtained. It is noted that very similar results are obtained by using the three Lyapunov functions. This can be explained by the size and simplicity of the power system which does not exhibit multi-generator instability.

Next, the fourth terms are included in the functions and the estimates for CCT's are recomputed as shown in Table 2. It is interesting to see that the additional term in the first two functions brings about an improvement in both CCT_1's and CCT_2's, the smallest errors being provided by V_2. Strangely, the inclusion of the fourth term in V_3 leads to relatively worse estimates.

Table 2. Critical Clearing Times for 4-Generator System

(a) Estimates of the Critical Clearing Times by using V_1 and V_{14}

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	132391[a]	0.557	0 502	-9.9	0 502	-9 9
2	172051	0 441	0 437	-0 9	0 394	-10.7
3	210030	0 491	0 489	-0 4	0 404	-17 7

(b) Estimates of the Critical Clearing Times by using V_2 and V_{24}

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	146499[a]	0 557	0 519	-6.8	0 519	-6.8
2	181546	0.441	0.443	0.5	0.406	-7 9
3	217816	0 491	0.488	-0 6	0 418	-14.9

(c) Estimates of the Critical Clearing Times by using V_3 and V_{34}

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	1 66349[a]	0 557	0.490	-12 0	0.490	-12 0
2	2.14772	0 441	0 428	-2 9	0 385	-12 7
3	2 71909	0 491	0.470	-4.3	0.383	-22.0

4.2 Results for 12-Generator System

Eleven type 1 points exist, one of which corresponds to the cluster of generators from 1 to 10 losing synchronism with the other two. The remaining ten points represent the generators, excluding 5 and 7, falling individually out of step with the rest of the system. The CCT estimates are shown in Table 3.

Table 3. Critical Clearing Times for 12-Generator System

(a) Estimates of the Critical Clearing Times by using V_1

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	161597[a]	0.117	0 113	-3 4	0.113	-3 4
2	903737	0.278	0.274	-1 4	0.138	-50 4
3	196096	0 301	0 347	15 3	0 326	8.3
4	808543	0.387	0.371	-4.1	0.207	-46.5
5	171611	0.287	0.142	-50.2	0.139	-51.7
6	828582	0 311	0 326	4.8	0.185	-40.5
7	171611	0 221	0.111	-49 8	0.108	-51.1
8	495784	0 324	0.342	5 6	0.236	-27.2
9	899148	0 300	0.302	0 7	0.144	-52 0
10	242594	0.401	0 295	-26.4	0.253	-36.9
11	6457812	0 278	0 318	14.3	0 063	-77.3
12	8348150	0 311	0 355	14.1	0 062	-80.1

(b) Estimates of the Critical Clearing Times by using V_2

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	394067	0.117	0 126	7 7	0 110	-6 0
2	1701419	0 278	0 303	9 0	0.127	-54 3
3	267375[a]	0 301	0 324	7 6	0 324	7 6
4	1262844	0.387	0 354	-8 5	0.163	-57 9
5	1752938	0.287	0 315	9.8	0.145	-49.5
6	1275566	0 311	0 342	10 0	0 185	-35.5
7	1752938	0 221	0.257	16 3	0.114	-48 4
8	792931	0 324	0.344	6 2	0.224	-30 9
9	1542023	0.300	0 304	1 3	0.126	-58 0
10	425643	0.401	0.270	-32.7	0.219	-45.4
11	5358124	0 278	0 304	9 4	0 083	-70 1
12	6499550	0 311	0.336	8 0	0 086	-72 3

A comparison of the estimates in Table 3 shows that with only one exception, local approximations again provide much more accurate estimates than the global one. As reported earlier [13], CCT_1's obtained by V_1 is considerably pessimistic and almost equal to CCT_2's for the cases of multi generator instability The use of V_2, however, not only eliminates this inefficiency, but also produces quite uniform and reasonable errors in CCT_1's (This uniformity is also observed to some extent in the analysis of the 4-generator system). For case 10, a reasonable estimate could not be obtained either by V_1 or V_2 There seems to be no apparent reason for such anomaly. The estimates computed by the method of hypersurfaces, for instances, are not worse that those for other cases It was not possible to use V_3 for the 12-generator system because it assumed negative values at all type 1 points Even though the function itself was found to be monotonically increasing in the neighborhood of the stable point, neither the closest nor the controlling u e p method was applicable

The effect of including the fourth term in each Lyapunov function, given in Table 4, is not positive, in general The most significant contribution of V_{14} is that the over-pessimistic CCT_1's are removed and replaced by highly accurate ones The term V_{24} does not seem to contribute much to the improvement of CCT_1's, in fact the uniformity of

results is a little lost Nevertheless, it is interesting to note the reduction of error in CCT_2's by about 5 percent on the average. This must normally be due to the inclusion of the effect of transfor conductances rendered by the term. The function V_3 together with its fourth term took on negative values at the type 1 points again, and thus could not be evaluated.

Table 4 Cntical Clearing Times for 4-Generator System

(a) Estimates of the Critical Clearing Times by using V_1 and V_{14}

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	240985	0.117	0.111	- 5.1	0 096	-17.9
2	919858	0 278	0.238	-14 4	0.102	-63.3
3	164018*	0.301	0 278	-7.6	0.278	-7.6
4	752599	0.387	0 292	-24.5	0.127	-67 2
5	1253517	0 287	0.283	-1.4	0 116	-59 6
6	733336	0 311	0.286	-8 0	0.152	-51.1
7	1253517	0.221	0.230	4.1	0.091	-58.8
8	444016	0 324	0.282	-13 0	0.185	-42 9
9	891269	0 300	0 245	-18 3	0.102	-66 0
10	255495	0 401	0.218	-45 0	0 177	-55 9
11	4168475	0.278	0.278	0 0	0 066	-76 3
12	5173594	0 311	0 310	-0 3	0.068	-78.1

(b) Estimates of the Critical Clearing Times by using V_2 and V_{24}

Faulted Bus	V_{cr}	Actual CCT (sec)	CCT_1 (sec)	Error (%)	CCT_2 (sec)	Error (%)
1	490252	0.117	0.132	12 8	0.112	-4.3
2	2117637	0 278	0 323	16 2	0.134	-51.8
3	307286*	0.301	0.335	11.3	0 335	11.3
4	1464617	0 387	0.369	-4.7	0.175	-54.8
5	2246332	0 287	0.342	19 2	0.154	-46 3
6	1386554	0 311	0.349	12.2	0.195	-37.3
7	2246332	0 221	0.283	28.1	0.122	-44.8
8	711069	0.324	0 329	1.5	0.236	-27 2
9	1387966	0 300	0.283	-5.7	0.133	-55 7
10	337940	0.401	0.242	-39.7	0.232	-42.1
11	5765358	0.278	0.311	11 9	0.089	-68 0
12	5326224	0 311	0 313	0 6	0 092	-70 4

Another interesting observation concerning Table 4 is the location of V_{min}. For V_1 this minimal value is obtained at the type 1 point associated with case 1 When V_{14} is added, or when V_2 or $V_2 + V_{24}$ is considered, however, V_{min} is computed at the type 1 point of case 3.

V. CONCLUSIONS

Three common Lyapunov functions are compared in the application of the closest and controlling u e.p. methods, which give global and local approximations to the separatrix. The results obtained using two test systems can be summarized as follows.

1) For the smaller test system, all the functions provide similar results. The inclusion of the forth term improves the estimates CCT_1's and CCT_2's.

2) For the larger test system, function V_1 produces over-pessimistic CCT_1's (which are approximately equal to the corresponding CCT_2's) in the cases of multi-generator instability Through function V_2, however, uniform results are obtained using local Lyapunov surfaces. The last function V_3 could not be used in the application.

3) When the functions ai corrected with the additional terms, no general improvement is gained. An important contribution of this correction in V_1 is the remarkable refinement of the estimates for cases where multi-generator instability occurs. Addition of the term V_{24} to V_2 does not improve CCT_1's in general, but reduces the error in most of CCT_2's. Function V_3 with V_{24} was again not applicable.

4) On the whole, V_2 is understood to be the most appropriate function among the ones examined here in generating local Lyapunov surfaces.

It is apparent that the performance of different Lyapunov functions in transient stability analysis of power systems varies quite considerably. This performance difference is exhibited especially when using the new approach, the controlling u e p. method. The major conclusion to be drawn out of this study is that although the use local Lyapunov surfaces decreases the conservativeness of the direct method of Lyapunov substantially, the controlling u.e p. method cannot always be employed as an effective and reliable tool as it is understood that its accuracy varies a great amount in different cases of instability occuring in power systems.

REFERENCES

[1] M. Rabbens-Pavella and F. J. Evans, "Direct Methods for Studying Dynamics of Large-Scale Electric Power Systems - A Survey", *Automat.ca*, Vol. 32, pp. 1-21, January 1985.

[2] A Bose, "Application of Direct Methods to Transient Stability Analysis of Power Systems", *IEEE Transactions on Power Apparatus and Systems*, Vol PAS-103, No 7, pp 1629-1636, July 1984.

[3] P. P. Varaiya, F. F. Wu and R.-L. Chen, "Direct Methods for Transient Stability Analysis of Power Systems. Recent Results", *Proceedings of the IEEE*, Vol. 73, No. 12, pp. 1703-1715, December 1985.

[4] H.-D Chiang, F. F. Wu and P. P. Varaiya, "Foundations of Direct Methods for Power System Transient Stability Analysis", *IEEE Transactions on Circuits and Systems*, Vol. CAS-34, No. 2, pp. 160-173, February 1987.

[5] T Athay, R. Podmore and S Virmani, "A Practical Method for Direct Analysis of Transient Stability", *IEEE Transactions on Power Apparatus and Systems*, Vol PAS-98, No. 2, pp 573 584, March/April 1979.

[6] M. Rabbens-Pavella, L T Grujic, J. Sabatel and A Bouffioux, "Direct Methods for Stability Analysis of Large Scale Power Systems", *Proceedings of IFAC Symposium on Computer Applications in Large Scale Power Systems*, New Delhi, India, August 16-18, Pergamon Press, UK , 1979.

[7] A. N. Michel, A A. Fouad, V Vittal, "Power System Transient Stability Using Individual Machine Energy Functions", *IEEE Transactions on Circuits and Systems*, Vol CAS-30, No 5, pp. 266-276, May 1983

[8] N. Kakimoto, Y Ohsawa and M. Havashi, " Transient Stability Analysis of Electric Power System via Lure Type Lyapunov Function", Parts I and II, *Trans IEE of Japan*, Vol 98, No. 5/6, May/June, 1978

[9] T Athay, V R Sherket, R Podmore, S Virmani and C Puech, "Transient Energy Stability Analysis", *Conference on System Engineering for Power*, Davos, Switzerland, 1979.

[10] A A Fouad, V Vittal, T K Oh, "Critical Energy for Direct Transient Stability Assessment of a Multimachine Power System", *IEEE Transactions on Power Apparatus and Systems*, Vol PAS-103, No 8, pp 2199-2206, August 1984

[11] H. Yee and B. D. Spalding, "Transient Stability Analysis of Multimachine Power Systems by the Method of Hyperplanes", *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-96, No. 1, pp. 276-284, January/February 1977.

[12] P. A Cook and A. M. Eskicioğlu, "Transient Stability Analysis of Electric Power Systems by the Method of Tangent Hypersurfaces", *IEE Proceedings*, Vol. 130, Pt. C, No. 4, pp. 183-193, July 1983.

[13] A. M. Eskicioğlu, "Determination of Transient Stability Regions in State Space Using Local Lyapunov Surfaces", *IEE Proceedings*, Vol 134, Pt. C, No. 3, pp. 234-237, May 1987.

[14] M. Ribbens-Pavella, "Critical Survey of Transient Stability Studies of Multimachine Power Systems by Lyapunov's Direct Method", *Proc. of the 9th Allerton Conference on Circuit and System Theory*, pp. 751-767, October 1971.

[15] J. L. Willems, "Direct Methods for Transient Stability Studies in Power System Analysis", *IEEE Transactions on Automotic Control*, Vol. AC-16, pp. 332-341, August 1971.

[16] A. H. El-Abiad and K. Napaggan, "Transient Stability Regions of Multimachine Power Systems", *IEEE Transaction on Power Apparatus and Systems*, Vol, PAS-85, pp. 169-179, February 1966.

[17] F. J. Evans, "Prospects for Dynamic Security Monitoring in Large Scale Electric Power Systems", *IFAC 7th World Congress*, Helsinki, Finland, 1978.

[18] M. El-Gwndi and M. Mansour, "Transient Stability of a Power system by the Lyapunov Method Considering the Transfer Conductances", *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-101, No. 5, pp. 1088-1094, May 1982.

[19] A. M. Eskicioğlu, "Transient Stability Analysis of Electric Power Systems by the Method of Tangent Hypersurfaces", *Ph. D. Thesis*, UMIST, Manchester, England, 1982.

MODELLING AND SIMULATION OF DIGITAL TRANSMISSION SYSTEMS: DESIGN OF OPTIMALLY TOLERANT EQUALIZERS[1]

R. T. VALADAS, Student Member, IEEE, R L.AGUIAR, Student Member, IEEE, A.M. de OLIVEIRA DUARTE, Member, IEEE

Integrated Broadband Communication Systems Group
Dept. Electronics and Telecommunications
University of Aveiro
3800 AVEIRO, PORTUGAL

Abstract - The design of a digital communication system is the result of some trade-offs between several system characteristics. For a well defined system an optimum design is usually difficult to achieve analytically. For that reason it is usual to resort to the help of appropriate software simulation and numerical optimization tools. The problem is further complicated if a complete characterization of the transmission channel is not known, as is the case when designing a system to operate in the customer access connection (CAC) or in the customer premises network (CPN). This paper presents a software package SPOT (*Simulation Package and Optimization Tool*), that calculates the optimally tolerant equalizer for a digital transmission system with unknown channel length in the absence of noise. In addition, a case study corresponding to the CPN environment is presented.

1 - Introduction

The equalizer of a digital transmission system is designed to compensate for the distortion introduced by a particular channel length. However, in most applications, such as the connection between a local exchange and a subscriber, the length of the channel is unknown whereas different subscribers can be at different distances from the local exchange. In this case it is usual to resort to the use of adaptive equalizers, but this can be a costly option. An alternative approach consists in using an optimally tolerant fixed equalizer.

This paper is concerned with the design of an equalizer optimally tolerant to channel length variations. In section 2, a brief description of the software tool used to simulate the digital transmission system is made. In section 3, the optimization problem is stated and the algorithm used is described. Some results concerning the CPN environment are presented in section 4 and in section 5 a brief summary is made.

2 - Simulation system

A software package SPOT (*Simulation Package and Optimization Tool*) is used to calculate the optimally tolerant equalizer for a digital transmission system with unknown channel length in the absence of noise. The model of the digital transmission system implemented by SPOT is presented in Fig. 1.

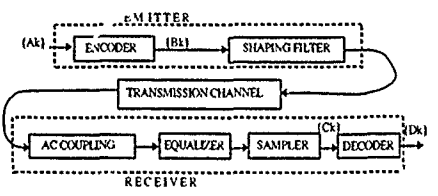


Figure 1 - Model of a digital transmission system

ENCODER. The encoder maps the input digital sequence $\{A_k\}$ into the transmitted digital sequence $\{B_k\}$. The line code can be chosen between the elements of a vast set which includes Manchester, CMI, 5B6B, NRZ, Walsh-2 and others.

SHAPING FILTER. The pulse shaping filter defines the format of the transmitted pulses. Two different shaping filters were included in our simulation system, a raised cosine with user-definable roll-off and an ideal low-pass filter.

TRANSMISSION CHANNEL + EQUALIZER. The transmission channel introduces distortion in the signal, hindering the correct decoding at the receiver. This distortion must be compensated by the equalizer.

Several types of transmission channel can be accommodated. Presently SPOT simulates two types of transmission channel, coaxial cable and optical fiber [1]. The characteristic parameters of the channel are normalized to the system bit rate. The models of the normalized frequency responses of the channel, H(f), and equalizer, E(f), are given by:

i) Coaxial cable:

$$H_c(f) = e^{-A_t\sqrt{2\,f\,(1+j)}} \quad \text{and} \quad E_c(f) = e^{A_e\sqrt{2\,f\,(1+j)}} \qquad (1)$$

ii) Optical Fiber:

$$H_o(f) = e^{-2(\pi A_t f)^2} \quad \text{and} \quad E_o(f) = e^{2(\pi A_e f)^2} \qquad (2)$$

where A_t is the channel distortion index and A_e is the equalizer index.

These equalizers are not realizable. Nevertheless, they are used as targets in most realization methods.

AC-COUPLING. An high-pass filter models the AC-coupling effects in the receiver, which can not be fully compensated by the equalizer. An nth order high-pass filter was implemented as a cascade of multiple first-order high-pass filters.

Two parameters, extracted from the eye diagram (ED), the aperture penalty P_o and the jitter penalty P_j, are used to assess the performance of the simulated system. The aperture penalty measures the decrease of the vertical eye opening introduced by intersymbol interference (ISI). The jitter penalty measures the unwanted pulse-position modulation of the recovered clock, introduced simultaneously by the shaping filter and ISI. Fig. 2 illustrates a typical eye diagram.

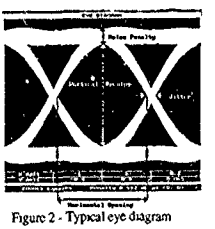


Figure 2 - Typical eye diagram

The aperture penalty P_o is given by

$$P_o = 20 \log_{10}(\theta) \qquad (3)$$

where θ is the worst-case normalized ED vertical opening. When $A_e = A_t$ and there are no AC-coupling effects, P_o equals 0 dB.

[1]This work was supported by RACE project R 1052 SPOT (*Signal Processing for Optical and Cordless Transmission*)

The jitter penalty P_j is given by:

$$P_j = 100.\left[\int_{-\frac{T}{2}}^{\frac{T}{2}} \left\{e[x]-m_e\right\}^2 p_e[x]\,dx\right]^{\frac{1}{2}} \qquad (4)$$

where $e(x)$ represents the "threshold crossing" instants in the ED, m_e is the average of $e(x)$, and $p_e(x)$ is the probability density function of $e(x)$. This corresponds to the r.m.s. of the signal transitions through a user definable threshold.

3-Optimization

3.1) Statement of the problem

The transmission channel length is assumed to vary between upper and lower limits, L_u and L_l, with a given probability density function $p(l)$.

In order to characterize the combined jitter and ISI distortion introduced by a particular channel length, a compound penalty measure is defined as

$$E_c = w\,P_0 + (1-w)P_j \qquad (5)$$

Here w is a weighting factor, describing the relative importance of jitter and ISI. This parameter is dependent on the actual hardware implementation. For example, a system with a degraded clock recovery circuit must be assigned a lower value of w. Usual values of w will be near 0.9.

A global error metric, the average compound penalty E_{ac}, is defined as the average of the compound penalties over the range of permissible channel lengths, i.e :

$$E_{ac}[A_e] = \frac{\left(\int_{L_l}^{L_u} E_c[l,A_e]\cdot p[l]\,dl\right)}{L_u - L_l} \qquad (6)$$

Definition:

The *optimally tolerant equalizer* is the one with an index A_e that minimizes the value E_{ac}.

3.2) Iterative Algorithm

At the onset of this work the main requirement for the numeric optimization algorithm was robustness. A particular variation of a constrained direct search method with random jumps [2] was developed. The iterative procedure is started with an heuristic first approximation and with a high step value. The algorithm searches a better approximation to the optimum equalizer using essentially the following rules:

i) Try the equalizer $A_e = A_{eo} + S$ (A_{eo} is the current choice of optimum equalizer and S is the current step value), and calculate $E_{ac}(A_e)$

ii) If A_e is a better equalizer than A_{eo}, then A_e is the new choice of optimum equalizer, S is reduced to S/2, and point i) is repeated.

iii) If A_e is not a better equalizer than A_{eo}, then step S is changed to -S/1.5, and point i) is repeated.

iv) Randomly (about once each three iterations) the step size is subject to a random change (about one fifth its present value), with the objective of preventing local minima.

v) The above rules are used with two exceptions. An acceptability threshold was established for preventing the design of an equalizer that would be in an average sense, or in a particular sense, a very bad one. If any E_{ac} or particular E_c exceeded a predefined value the equalizer was rejected (in these cases, either another line code should be tried, or adaptive equalization should be considered). Furthermore, the value of the optimally tolerant equalizer index was constrained to the range of channel distortion indexes defined by the user.

vi) The above rules are performed until a stop condition is reached. This could happen either by reaching a predefined accuracy or a predefined number of iterations.

vii) The iterative procedure (steps i to vi) is always run twice. This guarantees a further security against the occurrence of local minima.

The proposed algorithm performs in average as the classical 'Golden Rule' algorithm [5]. The interval reduction performed by the 'Golden Rule' is about 1.61. Although not oriented to an interval reduction approach, this algorithm performs a reduction between 1.5 and 2 in each iteration, with appropriated chosen starting step.

4 - Application to the CPN environment

As application example we choosed a situation taken from the work of one of the RACE projects concerned with the customer premises network: R.1035 CPN (Customer Premises Network). This project deals with the terminal to CPN interface in the future Integrated Broadband Communications Network (IBCN). The main assumptions are transmission using coaxial cable, 155.52 Mbit/seg bit rate, CMI line coding and the compliance with CCITT recommendation G.703 [4]. R.1035 considers a maximum characteristic attenuation of 12 dB, corresponding to a maximum length $L_u \equiv 150$m. A minimum length $L_l \equiv 10$ m is assumed, which corresponds approximately to a characteristic attenuation of 1 dB. R.1035 uses an 8 dB equalizer [3].

In the simulation process, the transmitted pulse shape was modelled as a 40% roll-off raised cosine. AC coupling, with a cut-off frequency of 150 KHz, was also considered. This value is clearly worst-case; even so its effects were found negligible. It was also assumed that channel lengths are *equiprobable*. Experimentally it was found that a simulation with N = 20 channels provided a reasonable resolution.

The results obtained are presented in Fig. 3. As it can be seen the optimum equalizer should present a characteristic attenuation parameter greater than 11.5 dB, if the jitter importance is reasonably lower than that of the ED vertical opening ($w > 0.8$).

The average compound penalties obtained with R.1035 8dB equalizer and our optimum equalizer are also depicted in Fig. 3. For w=0.9, it can be seen that the E_{ac} of the first one is 1.5 higher. This corresponds to an average 1.5 dB ED vertical opening penalty, if the two systems present no jitter penalty.

It is worth to mention that these results were obtained assuming an uniform distribution for the lengths of the transmission channel.

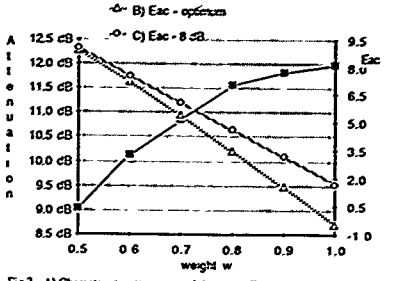


Fig 3 - A) Characteristic attenuation of the optimally tolerant equalizer, B) Average compound penalty associated with this equalizer and C) Average compound penalty associated with the equalizer used in [3], as a function of the weighting factor w

5 - Conclusion

A Simulation Package and Optimization Tool enabling the design of an equalizer optimally tolerant to channel lenght variations was presented. An application example related to the customer premises network environment was discussed. It was found that the use of an optimally tolerant equalizer can improve significantly the performance of a digital transmission system with a variable channel length.

6 - References

[1] "Comparative study of line coding, equalization and detection for optical fiber transmission in the Customer Premises Network", RACE Deliverable A1/UA/DA.2, RACE project R.1052/SPOT, January 1990.

[2] - "Non-Linear Optimization Techniques", M.J.Box, D.Davies, W.H.Swann, 1969, Oliver and Boyd, Ltd.

[3] "Preliminary Specification of the terminal to CPN interface", RACE Deliverable 35/FAT/TRS/DS/B/008/b3, RACE project R.1035/WP1, Customer Premises Network, July 1990

[4] Recommendation G 703, "Physical/Electrical Characteristics of Hierarchical Digital Interfaces", CCITT Blue Book Volume III, Fascicle III.4.

[5] "Optimization Methods", K.V. Mital, Wiley, 1976.

DATA SMOOTHING BASED ON L_1-MINIMIZATION AND ITS APPLICATIONS IN SIGNAL PROCESSING

L.N.Kotelnikova, N.G.Uzhakov
Institute of Microelectronics Technology and High Purity Materials, Academy of Sciences of the USSR,
Chernogolovka, Moscow District 142432, USSR

Abstract - The problem of nonparametric estimation of discontinuous functions is considered. In this case conventional methods are inefficient because they lead to the excessive smoothing near points of discontinuity. To solve the problem we suggest a special smoothing method based on L_1-minimization with following L_2-maximization. The method can have many applications in signal and image processing.

1. INTRODUCTION

Noise smoothing is a widely used operation in signal and image processing. The choice of the appropriate algorithm is determined by the kind of noise, by the class of processed signals and by the statement of the problem. In this paper the nonparametric estimation of discontinuous functions will be considered.

When the signal or image is described by a discontinuous function one often needs to recognize points (curves in two dimensional case) of discontinuity and to determine their positions and values of jumps as accurate as possible. In this case conventional methods are irefficient. We suggest a special smoothing method based on the using of L_1-minimization.

2. THE BASIC IDEA

The underlying idea of the method is as follows. Consider the regression problem where we have observations Y_i at points x_i, $i=1,2, \ldots, n$,

$$Y_i = f(x_i) + \xi_i ,$$

where f is the unknown function to be estimated, $\xi=(\xi_1, \xi_2, \ldots, \xi_n)$ is a vector of errors (for simplicity, we will assume that ξ_i are independent and identically distributed random variables with zero mean). If the distribution of ξ_i is known one can show positive α and β such that

$$P(-\alpha<\xi_i<\beta) > 1-\varepsilon$$

for any $\varepsilon>0$, or in other words,

$$P(Y_i-\beta<f(x_i)<Y_i+\alpha)>1-\varepsilon.$$

Assume, for simplicity, that ξ_i has a symmetric distribution. Then $\alpha=\beta$. Let us introduce a measure of roughness of any n-dimensional vector $b=(b_1, b_2, \ldots, b_n)$ as

$$\Omega_1(b) = \sum_{i=1}^{n-1} |b_i-b_{i+1}|$$

Consider the class Φ of all functions g such that 1) $Y_i-\alpha<g(x_i)<Y_i+\alpha$ and 2) g minimizes Ω_1 on the set of all functions satisfying to condition 1. It can be proved that under some conditions Φ consists of more than one element. Denote

$$\Omega_2(b) = \sum_{i=1}^{n-1} (b_i-b_{i+1})^2$$

Suppose the function $f(x)$ is estimated by $F(x,\varepsilon)$ for which the next two conditions are satisfied: 1) $F(x,\varepsilon)$ belongs to Φ, 2) $\Omega_2(F) = \max \Omega_2(g)$, where max is taken by all $g\in\Phi$.

The estimator F is very convenient for the estimation of discontinues functions. The first step (minimization of Ω_1) provides the smoothness (or more exactly nonroughness) of the estimate. The second step (maximization of Ω_2) prevents from the excessive smoothing near points of discontinuity. Note that the number α plays a role of smoothing parameter (see, for example, [1]).

The cube $[Y_1-\alpha,Y_1+\alpha]$, $i=1, 2, \ldots, n$, can be changed into the ball with the center $(Y_1, \ldots, Y_n)$ and some radius R (smoothing parameter).

Now let us assume that there is a sequence of finite sets:

$$X_1=(x_{11},\ldots,x_{1n_1}),\ X_2=(x_{21},\ldots,x_{2n_2}),\ \ldots,$$

such that

$X_{i-1}\subseteq X_i$, $i=1, 2, \ldots$, and their limit

$$\bigcup_{k=1}^{\infty} X_k$$

is dense on the function f domain of definition which assumed to be closed and without isolated points. Suppose that

$$P(|\xi_i|>t)=0(e^{-ct}),\ t\to\infty.$$

It is proved that under these conditions there exists a sequence $\alpha_1, \alpha_2, \ldots$, such that the sequence of estimators $F(x,\alpha_1), F(x,\alpha_2), \ldots$, is consistent for any point of continuity. If ξ_i are bounded then the sequence of estimators is strongly consistent (with probability 1).

3. CONCLUDING REMARKS AND SOME APPLICATIONS

In the previous section we only considered the simplest case. The method permits some extensions. First of all it can be used in the multidimensional case as well. Then the conditions for ξ (independence of ξ_i, their

identical distribution) can be essentially softened.

An advantage of the method is that it can be also used in the situation when the observed signal contains both the additive and multiplicative noise:

$$Y_i = (1+\gamma_i)f(x_i) + \xi_i ,$$

and besides the result is usually as good as in the case when there is only additive component.

Note that the suggested method can be used together with some other one by the follow scheme: on the first stage all points of discontinuity are searched by L_1-method, on the second stage the signal is filtrated by another method separately on each interval of continuity.

The method can have many applications in signal and image processing. In this section we consider in brief some of them.

A number of applications come from scanning electron microscopy. In microelectronics, for example, investigated objects often have structure which can be described by a piecewise constant functions of two or three variables. Besides, there are both of noise components: additive and multiplicative (multiplicative noise occurs as a result of beam current oscillations). In two dimensional case (surface investigation) a typical image usually consists of a number of simple figures like stripes, rectangles etc., which represent different values of some physical property (atomic number of the material, surface flatness etc.). The problem is to find boundaries of those figures. In three dimensional case images usually have more complicated nature but they are still described by discontinuous functions.

Another class of applications arises in spectrums processing. For example, the spectrum of electron losses is discontinuous (the point of discontinuity is related with the energy structure of electrons shells). Such spectrums are usually observed with a high noise so the point of discontinuity cannot be found directly.

Note also that the suggested method can be used for solving of so called change point problems.

REFERENCES

[1] D.M.Titterington, Common structure of smoothing techniques in statistics. International statistical review, 53 (1985), 2, pp. 141-170.

COMPUTATIONAL METHODS OF SCANNING ELECTRON MICROSCOPY SIGNAL PROCESSING

V.V.Aristov, A.V.Samsonovich, N.G.Ushakov, S.I.Zaitsev
Institute of Microelectronics Technology and High Purity Materials, Academy of Sciences of the USSR,
Chernogolovka, Moscow District 142432, USSR

Abstract - Mathematical aspects of signal processing of scanning electron microscopy (SEM) are considered. In particular it is investigated a class of integral equations describing the signal formation in SEM profilometry. Computational problems, relating with SEM study of multilayer structures, are also investigated. Models and algorithms of signal processing are presented for these cases.

1. INTRODUCTION

Scanning electron microscopy is one of the main diagnostic tools of semiconductor structures investigation. However there are some difficulties in its using for studying of submicron structures. Although the present day microscopes can provide a few nanometer resolution it is only possible for specially prepared thin samples with contrast inclusions. Examining the mass samples inherent in microelectronics a locality is determined by the dispersion zone size of primary electrons in a sample, and makes from fractions of micron to a few microns under different experimental conditions. Thus the signal has a complex integral nature and the validity of its interpretation requires use of special methods of signal treatment.

In this paper some problems of signal processing of scanning electron microscopy are considered.

2. SEM PROFILOMETRY SIGNAL PROCESSING

A number of SEM diagnostics problems leads to nonlinear integral equations. The first class of integral equations, we consider, comes from the problem of surface relief reconstruction (SEM profilometry). As it was shown in [1] the surface relief reconstruction is equivalent to the solving of an integral equation of the form

$$\iint_{-\infty}^{\infty} K(x,x_0,y,y_0)F(f(x,y)-f(x_0,y_0))dxdy = S(x_0,y_0)\ , \quad -\infty<x_0,y_0<\infty\ , \tag{1}$$

where $f(x,y)$ is the unknown function to be found (f describes the shape of the relief), K and F are two known functions satisfying to the following conditions: F is a continuous monotonic function, K is a positive continuous function such that

$$\iint_{-\infty}^{\infty} K(x,x_0,y,y_0)dxdy < \infty$$

for any x_0 and y_0, and

$$\iint_{-\infty}^{\infty} |K(x,x_0^*,y,y_0^*)-K(x,x_0,y,y_0)|dxdy \to 0$$

at $x_0^* \to x_0$ and $y_0^* \to y_0$

$S(x_0,y_0)$ is the observed signal. One of the main questions, related with the equation, is the uniquty of its solution. The following theorem shows that the solution of equation (1) is unique up to the addition of an arbitrary constant.

Theorem 1. Suppose that $f_1(x,y)$ and $f_2(x,y)$ are two solutions of equation (1). Then

$$f_1(x,y) = f_2(x,y) + \text{constant}.$$

Now consider the problem of the numerical solution of equation (1). It can be solved by the next iterative procedure:

$$\iint_{-\infty}^{\infty} K(x,x_0,y,y_0)F(f_{n-1}(x,y)-f_n(x_0,y_0))dxdy = S(x_0,y_0). \tag{2}$$

It takes place the next statement.

Theorem 2. Let $f_0(x,y)$ be an arbitrary initial estimate (any continuous function). Then the sequence $f_1, f_2, \ldots$, where f_n is determined from equation (2), converges in the uniform norm to the solution of equation (1).

3. SIGNAL PROCESSING OF SEM DIAGNOSTICS OF MULTILAYER STRUCTURES

The investigation of multilayer structures is a very important problem of SEM diagnostics, especially in microelectronics. However, until recently the most important types of signal (secondary and backscattered electrons, electron beam induced current etc.) have been used only for surface analysis. The main reason of this is the intricate dependence of the generated signal on the object structure in the case of multilayer specimens investigation. In this section of the paper we discuss matters connected with models and algorithms of signal processing arising in multilayer structure diagnostics.

Assume that the object in question has an ideally smooth surface and inhomogeneous distribution of the atomic number $f=f(x,y,z)$ within itself. f is an unknown function which should be reconstructed. Let the electron beam be incident to a point (x_0,y_0). Denote by E the electron beam energy and by $P(x_0,y_0,E)$ the secondary electron emission coefficient. Then the signal formation can be described by the following equation

$$\iiint K(x-x_0,y-y_0,z,E)R(z/E^{\gamma})f(x,y,z)dxdydz = P(x_0,y_0,E)\ , \quad -\infty < x_0,y_0 < \infty\ ,\ \gamma>0, \tag{3}$$

where K and R are known functions. If the electron beam energy E is fixed, then equation (3) is underdefined: the unknown function depends on three variables whereas the rihgt-hand side only on two. Hence, f cannot be unambiguosly reconstructed by P. To make equation (3) unambiguosly solvable for f, the energy E should be made variable. Equation (3) is a special case of the Fredholm equa-

tion of the first kind and represents an ill-posed problem. If the energy interval is wide enough then it can be solved numerically by a special regularization method based on use of the Fourier or Mellin transform.

Another class of numerical problems arises in electron beam induced concentration methods [2].

The methods of induced concentration (MIC) include the following methods: EBIC, TREBIC, IRBIC, EBIV, SHF-absorption, cathodoluminiscence, therroacaustics, SDLTS. The measured contrast c in all these methods is spatial-dependent if timelife τ of induced carriers is not uniform, $\tau(x,y,z)$

$$c(r_0)=[(I_0-I(r_0,(\tau)))]/I_0, \quad r_0-(x_0,z_0)$$

Of course the information (two-dimensional function $c(x_0,z_0)$) is not sufficient for reconstruction three-dimensional function τ (solving the inverse problem). We suggest to use as a third coordinate energy of incident beam E_B. The equation of signal is

$$c(r_0,E_B)=\frac{1}{I}\int d^3r\ Q(r)\ \Gamma(r)\ P(r,r_0,E_B), \qquad (4)$$

where concentration of minority charge carriers, P(r), is a solution of the diffusion equation

$$(-D(\nabla+(\nabla\phi))+1/\tau_0+\Gamma)P=g,$$

$$\Gamma(r)=1/\tau(r)-1/\tau_0$$

The signal equation is valid for any method mentioned above because of the factor Q is known function for each case. The dependence of P on life time τ makes equation (4) nonlinear. We proved the nonlinear problem (4) is equivalent to linear equation for renormalized function Γ_{eff} and simple relation between Γ and Γ_{eff}

$$c(r_0,E_B)=I_0^{-1}\int d^3r\ Q(r)\ \Gamma(r)\ P_0(r,r_0,E_B),$$

$$\Gamma(r)=Q(r)\Gamma(r)/[Q(r)-$$

$$\int d\ r'Q(r')\Gamma(r')G(r',r)]$$

Here P_0 and G are solution and Green-function of diffusion equation for uniform case. Now we can say that well known Donolate approximation gives namely Γ_{eff} but not the real time-life. Of course, if non-uniformity of τ (the value of contrast) is small the functions Γ and Γ_{eff} are very close. The computer simulation has shown that it is necessary to measure contrast with high accuracy (relative error 10^{-3}-10^{-4}).

In a number of situations an analytical model of the signal generation is not available. For such cases we suggest the use of Monte Carlo simulation as a tool to solve the inverse problem.

REFERENCES

[1] V.V.Aristov, V.V.Kazmiruk, N.G.Ushakov and Firsova, Poverkhnost 4(1989), 120 (in russian).

[2] S.I.Zaitsev, A.B.Sansonovich, Izvestija AN SSSR, 54(1990), 2, 247 (in russian).

TOWARDS QUALITATIVE CONTROL ENGINEERING.

Féray-Beaumont S.*
Process Control lab.
ICPI-Lyon; 31, place Bellecour
69288 Lyon cedex 02, France.

Morris A.J.
Department of Chemical and Process Engineering
University of Newcastle; Merz court, Claremont road.
Newcastle upon Tyne NE1 7RU, UK.

Abstract: Unlike the local loop control strategies which are now usually performed by numerical algorithms on many process plants, the supervisory level may still be devolved to human operators. Recently developed techniques in Qualitative Reasoning may be useful to aid operators within some of their tasks. Initial results in Qualitative Physics showed encouraging potential. However, such an approach may not be the most appropriate basis from which to start in addressing process supervision problems. Control engineering principles, encoded in Qualitative Transfer Function form, seem to provide a more widely applicable technique.

1 Introduction

One major irony of increasing process plant automation is the demand for improvements in operators skills and the need for more effective training in order to enable them to better carry out their supervisory tasks [Bainbridge83]. Supervisory control implies a global understanding of the process evolution which is be difficult to capture in mechanistic model, especially in complex plants. Moreover the results provided by numerical simulations may not be at the desired level of abstraction for supervisory tasks.

There are therefore motivations to develop a qualitative model-based reasoning technique and recent work on qualitative modelling and simulation have shown the validity and the relevancy of such approaches. Initial results have been encouraging and steps are now being taken to apply Qualitative Reasoning methods to real world problems. Whilst a deep knowledge of the physical characteristics of a process, providing they can be expressed, might be useful for control purposes, Qualitative Physics [deKleer84] [Kuipers89] may not provide the most useful level of abstraction to deal with process supervisory issues. Mapping the problem into control engineering transfer function form and adopting qualitative modelling concepts could lead to a "qualitative control engineering" approach obviating the well known ambiguity problems in Qualitative Physics modelling.

2 Qualitative Control Engineering?

It is now broadly agreed that Qualitative Reasoning has a major role to play in high level supervisory control on process plant [Kuipers89] [Féray91] [Leitch89]. However, the requirements for supervisory control are different from those for closed loop automatic control. Loop control, optimisation, etc., is based around numerical models of appropriate structure and detail. In many situations overall process supervisory control tends to be based around human operators who make use of both qualitative descriptions of approximate relations between variables and direct quantitative measurements. Different goals, different techniques, different conditions of application demand different tools. In order to drive supervisory control closer to automatic control philosophies, Qualitative Reasoning developers have attempted to mimic physicists' reasoning and mathematics - creating so called deep knowledge. Control engineering philosophies seem to be a more appropriate basis upon which to build process plant supervision strategies.

Some initial steps have already been made. Qualitative techniques have been used to reveal some structural properties of physical systems - controllability, stability, etc. [Travé86]. Work is still being carried out, however, on qualitative algebras [Missler89] in order to provide Qualitative Reasoning with a systems related formalism. Concepts such as qualitative matrix, qualitative rank of a matrix, qualitative eigenvalues, etc., have been defined with a view to transposing the results of "classical" control engineering into their qualitative counterparts (e.g. pole assignment).

Another approach consists of providing a view of the relationship between two variables, in terms of input and output, without knowing exactly the physical relationship between the two variables. This corresponds to the methods adopted in Transfer Function theory. Thus, a qualitative counterpart would be to describe the evolution of related variables through behaviour constraints qualitatively described.

3 Qualitative Transfer Function-based modelling.

3.1 Causal Graph and Qualitative Transfer Functions.

A major advantage of Transfer Function methodologies is the simplicity of representation A similar expressive form is achieved by using a causal graph model of the process [Feray89], instead of a set of differential equations (numerical or qualitative [deKleer84]) to describe physical phenomena Each node of the graph represents a variable relevant to that required for process supervision purposes. The arcs are described as the behaviour constraints (called Qualitative Transfer Functions- QTF) between the variables. The QTF-based model encodes the concepts used within the Transfer Function (TF) framework. The resulting qualitative model is then to be used to simulate process behaviour in order to provide a means of early detection of process maloperation, with subsequent fault diagnosis.

The evolution of a variable (called *history*) is a piece-wise linear function linking the important steps in the variable behaviour. These "steps" consist of remarkable points on the evolution of the process variables (e.g. alarm threshold overshoot) Between any two points (called *events*), the approximation is linear for the purposes of simplicity. QTFs are defined as the way of relating two piece-wise linear evolutions. They are approximations of the a-priori well-known responses of numerical TFs to standard inputs (steps and ramps) For consistency, these responses are transformed into piece-wise linear functions which can be seen as sets of intervals parameterised by the values of the temporal variables describing the QTF (time-delay, settling time or period of oscillation).

In order to describe the different forms of behaviour constraints between continuous variables, a QTF

* To whom all correspondence should be sent.

library is proposed [Féray89]. The QTF library has an open-ended structure. Behaviour constraints can always be added if needed (e.g. integrating and non-minimum phase characteristics, etc.). Some other kinds of propagation constraints including fuzzy functions are proposed in [Vescovi90].

3.2 Simulation Algorithm.

The qualitative simulation is based on an event-based propagation algorithm [Leyval90]. This algorithm is asynchronous and fed by data measured from the process (or from any other source). The propagation through the graph is achieved by transforming all events, sorted according to the time, into a set of events using the QTFs and merging these newly created events into the existing histories of the variables. The number of new events in a set, their "value" and "time", depend on the input event and on the information contained in the arc (i.e the kind of behaviour constraint, and the different parameters). It is terminated when there is no event left to be propagated or when the propagation has been completed within a pre-determined time interval (in case of loops in the graph).
The relevant information produced by a QTF-based model, is the shape of the behaviours, the orders of magnitude, etc. Such information is sometimes more useful for process supervision, fault diagnosis, etc. than precise numerical values [Leitch89].

4 Applications.

The first application of the QTF approach was for the modelling of a steam generator of a nuclear power plant [MQ&D91]. The aim of the system was to provide the process operators with some useful information to help them prevent unwanted and costly shut-downs of the steam generator.
Another application of the QTF approach to a different kind of plants, pulsed columns, is described in [Féray89]. Work is now being carried out to provide an analysis of both qualitative simulation and the measurements on the process in order to achieve a better understanding of the process evolution [Montmain90]. It is reported that another application involving the same system will start in 1991 on a nuclear fuel treatment plant [MQ&D91].
The general modelling tools and the simulation algorithm have also been used to build a qualitative model of a distillation process [Feray91]. Using a complex numerical physical-chemical model as a reference, the results produced by the QTF-based model are compared with those from the detailed numerical model. The relevancy of the information provided is then able to be assessed. Comparisons have been made on both the top and bottom product flows and both the top product and product compositions, and demonstrates the validity of the approach [Feray91].

5 Discussion:

The motivation for our work is to build simple models encoding both approximate "behavioural" knowledge and more numerically accurate data, when available, within the structured framework of a QTF-based model. The qualitative simulation results are mainly composed of evolution tendencies of the process variables and their orders of magnitude.
It is stressed that qualitative modelling does not aim at replacing numerical modelling in control loops, but at complementing the existing loop control level with a high level decision making loop providing information that is closer to the human operators' way of reasoning.
Qualitative Control Engineering provides a set of tools for process supervision built on the well known classical and modern theories for process control. Such an approach appears to provide the prospects for a major step forward in Qualitative Reasoning. If some of the statements made by MacDermott in his article "Artificial Intelligence meets Natural Stupidity" bear fruit, then Qualitative Control Engineering could, in the short term, aid operators in their supervision task whilst ultimately contributing to the goal of a totally computer based control strategy.

References.

[Bainbridge83] L. Bainbridge. Ironies of automation. Automatica vol. 19. n 3. 1983.

[deKleer84] J. de Kleer, J.S. Brown. A Qualitative Physics based on confluences. Artificial Intelligence vol. 24. 1984.

[Féray89] S. Féray Beaumont, L. Leyval, S. Gentil. Declarative modelling for process supervision. IFAC/AIPAC Nancy (France). 1989.

[Féray91] S. Féray Beaumont, M.T. Tham, A.J. Morris. Towards distillation process supervision using Qualitative Model-Based Reasoning. ACC '91. Boston. 1991.

[Kuipers89] B. Kuipers. Qualitative Reasoning: Modeling and Simulation with incomplete knowledge. Automatica. Vol. 25. n 4. 1989.

[Leitch89] R.R. Leitch, M.E. Wiegand, C. Quek. Coping with complexity in physical system modelling. Third International Qualitative Physics Workshop. California. August 1989.

[Leyval90] L. Leyval, S. Féray Beaumont, S. Gentil Event oriented versus interval oriented qualitative simulation. Mathematical and Intelligent Models in System Simulation. IMACS-IFAC. Brussels (Belgium). 1990.

[MacDermott77] D. MacDermott. Artificial Intelligence meets Natural Stupidity in Mind Design. J. Haugeland ed. Bradford Book. 1977.

[Missier89] A. Missier, N. Piera, L. Travé-Massuyès. Order of magnitude qualitative algebras: a survey Revue d'Intelligence Artificielle, Vol. 3, No 4, 1989.

[Montmain90] J. Montmain, L. Leyval, S Gentil On line qualitative interpretation of dynamic simulation for diagnosis. Mathematical and Intelligent Models in System Simulation IMACS-IFAC. Brussels (Belgium). 1990.

[MQ&D91] MQ&D Project Qualitative Reasoning: Methods, Tools and Applications. Editing coordinator: L. Travé-Massuyès. LAAS Toulouse (France). 1990.

[Travé86] L. Travé, E. Kaszkurewicz. Qualitative Controllability and Observability of linear dynamic systems. IFAC/IFORS symposium in Large Scale Systems, Zurich, Switzerland. 1986.

[Vescovi90] M. Vescovi A new model for qualitative reasoning representing time explicitly Mathematical and Intelligent Models in System Simulation. IMACS-IFAC. Brussels (Belgium) 1990

INTERPRETATION BASED ON QUALITATIVE REASONING CONCEPTS

FRANCOIS GUERRIN
Institut National de la Recherche Agronomique
Artificial Intelligence Laboratory
B.P. 27 - 31326 Castanet-Tolosan (France)

Abstract - This paper deals with Qualitative Reasoning concepts for simulating the interpretation of measurements and observations performed on complex systems for management purposes. Starting from the definitions of interpretation, explanation and causality, the continuity existing between these three basic notions is shown, and a (preliminary) theory of interpretation is proposed. Then, the methodology used for designing a model of interpretation applied to the field of biological processes is described.

I. INTRODUCTION

In the domain of system supervision (Caloud, 1988; Feray Beaumont et al., 1989), the concept of interpretation seems to be particularly important, as continuous processes internal to complex systems are not directly observable. They can only be apprehended by the use of sensors, analyses or observations, i.e. in a discrete and external way. Data that are collected on them constitute the set of signs discerned by the observer. If eventually they can have a direct meaning (are not they themselves an interpretation of real phenomena ?), they get only their full signification through interpretation, itself related to the main goal to be reached.

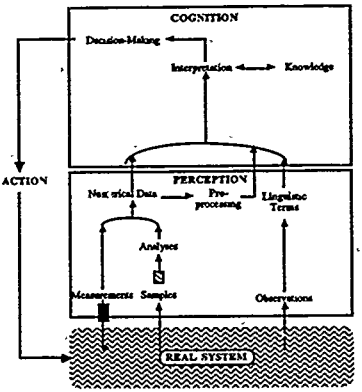


Figure 1 - The control loop for system management.

Interpretation is obviously based on subjective estimations, because external world signals, received by our senses, are interpreted through our perceptive categories, molded by self-experience (Ninio, 1989). Interpretation is essential, both in the constitution of operator's knowledge and in its ability to operate. Interpretation prepares decision-making, and thus appears to be the first step of Cognition, central process in the control loop, just between Perception (i.e., data collection) and Action (Fig. 1).

In fact, a human operator behaves as an Expert (Vogel,1988). Judgment and action rules, progressively constructed, are applied to data (eventually preliminary processed), usually translated into qualitative terms to become the basis of reasoning (De Kleer and Brown, 1983). This high level approach is by nature action-oriented, as it integrates a global knowledge simplifying the real world complexity to make it understandable and controllable (Rasmussen, 1985). Hence, its representation can to some extent fill the gaps in mathematical modeling, when no acceptable model is available or when solutions generated must themselves be interpreted before being used in decision making (Caloud, 1988). We believe therefore that it does exist a space for a complementary approach, aimed at computing all available information on the system, including the pieces which would be too difficult to formalize in terms of classical mathematical equations.

II. A THEORY OF INTERPRETATION

But what is interpretation ? According to Ferida (1980), the "art of interpretation" is based on the recognition of a sense, hidden under the apparent meaning taken by god's word, the manifestation of a sign, the expression of a gesture or a word... In agreement with that definition, giving an interpretation is an action of sense production by means of a discourse applied to signs. This general concept is applied to such various fields as theological exegesis, art, psychoanalysis, human sciences, medicine...

How about explanation ? According to Varela (1989), explaining is reformulating phenomena in such a way that their components be causally related to each others. Thus, if interpretation is seen as a discourse about things, explanation can be seen also as a part of interpretation.

How causality takes place within these concepts ? In agreement with Saint-Sernin (1980), we will start from the idea that the term causality does not mean directly a property of the relationships between objects (ontological meaning), but a characteristic of the statements constituting an explicative theory of observed facts. Therefore, if explanation is considered like the way to establish causality, and causality like the way to explain facts, there is a clear equivalence between these two notions.

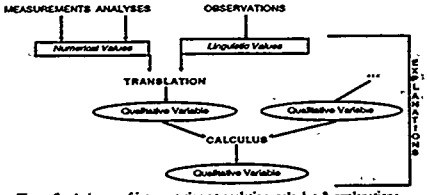


Figure 2 - A theory of interpretation : translation, calculus & explanations.

Finally, our "theory" of interpretation, considered as meaning assignment to observed signs, includes three basic levels necessary for the comprehension of the system under observation (Fig. 2) :

- *Translation* : assignment of individual meaning to the model input-values (e.g., the measured values x, y, z... of variables A, B, C will be qualified as low, bad, slow..., according to their nature);
- *Calculus* : assignment of individual meaning to all the model internal variables (e.g., combining several qualitative values A = low, B = bad , C = slow..., will allow one to deduce the qualitative values K = good, L = high..., of unmeasurable variables K, L...);
- *Explanation* : assignment of meaning to the system, by enunciating all the causal relationships between variables accounted for translation and calculus (e.g , A = low and B = bad imply K = good...).

Thus, giving an interpretation is deducing, from any subset of input-values, qualitative values of as many unknown variables as possible, and explaining the reasoning to provide the user with an overall comprehension of the phenomena.

III. METHODOLOGY OF INTERPRETATION

This approach was applied to data interpretation in hydro-ecology, i.e. interpretation of measurements, analyses and observations performed on an aquatic ecosystem for management purposes (Guerrin, 1990; Guerrin, 1991). It involves four methodological steps for designing an interpretation model based on expert knowledge (skilled process engineer's) :

1 - Identification of relevant variables for system supervision purposes, especially unmeasurable or unobservable ones;

2 - Representation of influences between the variables as a cause-effect graph, by starting from the identification of elementary causal chains;

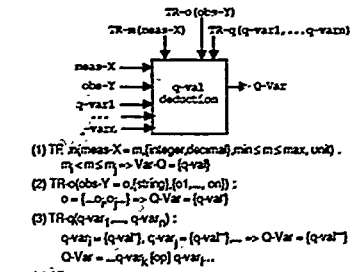


Figure 3 - Specification framework of qualitative transfer rules (TR-m . translation of measurements; TR-o . translation of observations, TR-q : calculus, AR . action rules).

3 - Knowledge specification, according to a standardized framework which distinguishes (Fig. 3) :

• Translation Rules (*TR-m* and *TR-o*) of numerical data (measurements, analyses) and linguistic terms (observations) respectively, into a relevant vocabulary for the application : e.g., symbols like {pp, p, m, f, ff} whose semantics corresponds to orders of magnitude (e.g., very low, low, medium, high, very high, respectively);

• Calculus Rules (*TR-q*) operating on this set of symbols (called Quantity Space) to deduce the qualitative value of a variable from qualitative values of its causes, by using 6 formal calculus operators (definitions given Fig. 4 below);

• Action Rules (AR) for controlling the use of above-cited rules.

4 - Software programming (that was made with PROLOG); at this level, giving an interpretation consists essentially in producing, from input-values (measurements, observations, given by the user or delivered by sensors) :

• the system State . i.e , the set of all the generated estimations in the form of <qualitative-value, list of qualitative variables>,

• Explanations : i.e., in agreement with De Kleer and Brown (1983), the execution trace of reasoning (at least) having led to the calculus of variables throughout all the paths available in the causal graph, displayed in a language that can be understood by the user.

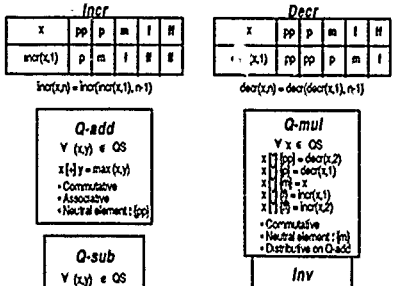


Figure 4 - Definition of six operators for qualitative calculus on a five symbols space QS = {pp, p, m, f, ff}.

IV - DISCUSSION

We must point out that our system performs a static reasoning, as it gives an interpretation of a set of measurements, analyses and observations made at a particular instant. As it works in a deductive sense, it is prediction-oriented (given the causes determine the effect). But these principles of interpretation could also be applied to the diagnostic sense (given the effect determine the causes). Beyond this aspect, we believe that taking into account the past evolution of variables (i.e., interpretation of the succession of memorized states) would be of major interest. Integration in qualitative reasoning of information extracted from time evolution curves, seems therefore fundamental (Willoughby, 1989). We are beginning to address this topic which highlights some interesting problems : recognition of patterns, intelligent smoothing, kinetic comparison...

Our formal calculus system has a certain lack of generality, for example : m [+] p [+] p [+] p... will always result in {m}. Calculus are therefore restricted to relatively shallow reasoning. On the other hand, Quantity Space symbols represent actually adjacent intervals. Fuzzy intervals would be perhaps, conceptually better. This was thought unnecessary because in our application, adjacent intervals fit the expert's way of thinking, without generating spurious results. Two other problems must be emphasized, although they appear more specific to living-systems (Varela, 1989) :

- first, is the problem of "circular causality", when an effect and its cause are confounded; we dealt with it by choosing an unidirectional sense of causality according to our purposes (i.e , by breaking the loop) ; we hope that better solutions can be found,
- then is the problem of designing a real dynamic model, because of a lack of knowledge about temporal relationships between variables such as delay or settling time.

Our actual perspective (outside the scope of time evolution curve analysis) is mainly to apply these concepts to other biological systems We are beginning a work on fermentation reactors Our preliminary experience in that field is encouraging. Beyond this, a longer-term perspective would be trying to build something like a "General Theory of Interpretation", valid at least in the field of biological processes...

REFERENCES

Caloud P , 1988 - Raisonnement qualitatif, application à l'aide à la supervision des procédés continus. Ph.D. thesis, INPG & LIFIA, Grenoble, 150 pp.

De Kleer J., Brown J.S., 1983 - The origin, form and logic of qualitative physical laws. In IJCAI 83, 8-12 Aug , Karlsruhe, 2, pp. 1158-1169.

Fedida P., 1980 - Interprétation. In : Encyclopaedia Universalis, vol. 9, pp. 30-33.

Feray Beaumont S., Gentil S., Leyval L., 1989 - Declarative modelling for process supervision. Revue d'Intelligence Artificielle, 3 (4) : 135-150.

Guerrin F., 1990 - Valorisation aquacole d'eaux usées traitées par lagunage naturel, évaluation biotechnique et modélisation des connaissances. Ph.D. thesis, P. Sabatier Univ., Toulouse, 297 pp.

Guerrin F., 1991 - Qualitative reasoning about an ecological process interpretation in Hydro-Ecology. Ecological Modelling, (accepted January 1991).

Ninio J., 1989 - L'empreinte des sens. Odile Jacob, Paris, 265 pp.

Rasmussen J., 1985 - The role of hierarchical knowledge representation in decisionmaking and system management. In : IEEE transactions on systems, man and cybernetics, smc-15, n°2, pp. 234-243.

Saint-Sernin B., 1980 - Causalité. In : Encyclopaedia Universalis, vol. 3, pp. 1089-1092.

Varela F.J., 1989 - Autonomie et connaissance; essai sur le vivant. Seuil, Paris, 254 pp.

Vogel C., 1988 - Génie cognitif. Masson , Paris, 196 pp.

Willoughby A., 1989 - Making qualitative reasoning more quantitative. In : Avignon 89, Second Generation Expert Systems, 29 may-2 june, pp.117-127.

QUALITATIVE FAULT DETECTION VIA FUZZY ANALYSIS

Jacky MONTMAIN
Commissariat à l'Energie Atomique
CEN.VALRHO - Marcoule
DPR/SCD
Laboratoire d'Informatique Appliquée
BP 171
30205 Bagnols-sur-Cèze Cedex
Phone : 66 79 66 31

Sylviane GENTIL
LAG-ENSIEG
UA CNRS 228
BP 46
38402 St Martin d'Hères Cedex
Phone : 76 82 63 85

ABSTRACT

This paper deals with a qualitative interpretation of an on line simulation for supervision. The simulation algorithm relies on the propagation of events through a causal graph : each signal is decomposed as a series of significant variations. The fault detection rests on a qualitative interpretation of the difference between the process and the simulation. This comparison has to be robust enough to avoid the classical drawbacks of the threshold methods and to take into account the disturbances inherent to the modelling or to the measurement system.
Firstly modelling techniques are briefly reported and the notion of significant event in a process is presented. Next, several specific terms are introduced to provide a definition of qualitatively similar behaviour. Finally, the algorithm describing the qualitative interpretation of the error is presented . it is based on fuzzy analysis.

KEY WORDS

Qualitative interpretation - Supervision - Fault detection - Diagnosis

1. INTRODUCTION

DIAPASON is a system which aims to assist operators in supervising industrial continuous processes. It provides operators with the simulation of the normal behaviour of the process and with a help for high level diagnosis [1]. It is a modular system . PROTEE is the qualitative simulator whose model is a causal graph [2] , MINOS analyses the simulation results by comparing them on line with the process measurements, and starts a fault diagnosis when necessary [3] , SPHYNX diagnoses by providing the causes of the detected defect thanks to expert rules that use a structural knowledge of the process [4].
This paper deals with the control and decision unit, MINOS, and describes more especially the comparison algorithm. The aim is to detect complete and drift failures of the process as soon as possible thanks to model-process comparison.
An image of the normal behaviour of the process is obtained owing to the simulator PROTEE. The model of the process is expressed as a graph whose nodes are the variables relevant to the operator and whose arcs represent the causal relations between the variables. This model is not based on a sound physical analysis of the various phenomena occuring in the process. The evolution of a variable is represented by a piecewise linear function. The simulation algorithm takes into account step or ramp evolutions according to dynamics of signals. An evolution is propagated through an arc piece after piece ; the response is still a piecewise linear time function with discontinuities, named Qualitative Response
The signals to be compared are the measured signal which is sampled and the simulated one, a Qualitative Response. In order to have regard to homogeneity, the sampled measured signal has to be replaced by a segmented representation, it means by a piecewise linear time function. This approximation is not very restrictive concerning the calculated error . the use of the sampled signal of the process would have provided an illusory precision with regard to the point of view of the modelling and the quality of industrial measurements. Moreover, this substitution is beneficial to the storage of the histories of variables . it is far less costly to store series of significant events than periodically sampled values.
The simulation results correspond to a synthetic representation of the process and to the level of abstraction relevant to supervision. From this point of view, the shape of an evolution is as important as the precise values of its points. So it is necessary to use integral criteria on temporal windows rather than instantaneous criteria in order to compare the simulated signals and the real ones.
The causal model also used for diagnosis has a numerical parameterization : gains, delays, settling times ... But the nature of the simulated signals is rather qualitative . approximation of integral signals by piecewise linear functions.
The model-process comparison is the basis of the decision unit MINOS which has to start a finer diagnosis when necessary. Starting a diagnosis on a simple comparison of the simulation error with a threshold makes a system generally efficient in the detection of defects but too sensitive to the various disturbances and imprecisions (it leads to bad detections). To avoid this problem, qualitative notions which are more consistent with the modelling principles are introduced in MINOS ; the algorithm rests on a local treatment on each simulated variable and takes into account the various sources of disturbances ; this provides therefore an algorithm robust in dynamic phases.

2. DEFINITIONS AND TERMINOLOGY

An event e_i is a significant change in the behaviour of a variable, it means a variation able to modify soundly the global behaviour of the process , it is parameterized by its initial time, its amplitude variation and its slope variation. It is noted . $(t_i, \Delta p, \Delta a)$ When the next event $(i + 1)$ occurs, the duration of event i can be calculated . $t_{i+1} - t_i$
An **evolution** is a sequence of events whose durations are not equal The response to an evolution through an arc of the causal graph is still an evolution (stability of the representation for the simulation algorithm). An evolution is represented by a piecewise linear discontinuous time function which is deduced by an elementary bijection. From evolution [(0 0, 1 0, 0 0) , (1.0, -1 0, 0 0) , (3.0, 0 5, 0.5)], the piecewise linear time function can be easily deduced in figure 1 :

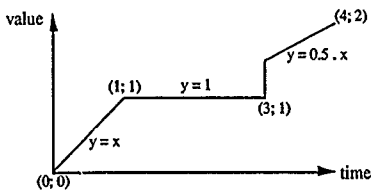


Figure 1 : An evolution

The simulation is started by significant events detected on the control variables or on measurable disturbances of the process. An event detection algorithm is needed. After filtering noise, the detection must filter measured evolutions so as to exclude too small amplitude or too short duration events. This algorithm which runs on line is called segmentation . it enables to transform periodically sampled

signals into evolutions. Two main parameters are necessary for adjusting the segmentation according to dynamics of the evolutions : an amplitude threshold and a temporal window width.
A temporal window TEMPO is defined for each variable so as to analyse the behaviour on a significant duration.

A basic error criterion ε(t) is defined for the comparison : it is a signed integral error between the simulated evolution x*(t) and the segmented measured one x(t) :

$$\varepsilon(t) = \frac{1}{TEMPO} \cdot \int_{ti}^{ti+TEMPO} (x(t) - x^*(t))\, dt \cdot \qquad (1).$$

This allows fluctuations of the simulated evolution around the measured evolution on the window TEMPO. The cost of the calculus ε(t) is very low for it corresponds only to an integration of piecewise linear time function.
Describing qualitatively a behaviour has lead us to extend the notion of similar evolutions : the use of a moving temporal window enables to define a temporal average, a standard deviation, and more generally it gives access to a local analysis of the compared evolutions over the selected window (tendencies and shapes...).
A tendency is a symbolic scale relative to the derivative of a function ; three levels are distinguished : INCREASING, DECREASING and STEADY. A tendency is the result of the analysis on a temporal window.
The evolution of a variable is available on a temporal window TEMPO at every moment. Then it is possible to get the corresponding simulated history of the variable on a second temporal window whose width would be : TEMPO + 2.0 * MARGIN. MARGIN is named the temporal margin and it enables to have the simulated evolution on a larger duration on both sides of the measured one at one's disposal (figure 2) :

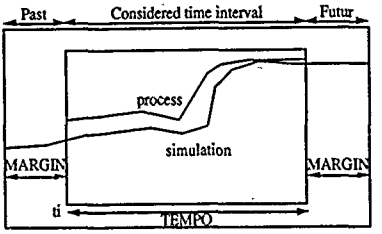


Figure 2 : Defined temporal windows

So translations in the plane (time, value) make it possible to search for the part of the simulated evolution, on a duration TEMPO, extracted from the complete simulated evolution, defined on [ti - MARGIN ; ti + MARGIN + TEMPO], which approximates as well as possible the corresponding measured evolution on [ti , ti + TEMPO]. The translation in the plane that provides the minimization of the quadratic deviation between the simulated evolution and the measured one on [ti ; ti + TEMPO] corresponds to the optimal proximity (OP) of the real behaviour [5]. If the quadratic deviation between both evolutions provided by the optimal proximity is below a given threshold, both evolutions have similar shapes.
The associated criterion to be evaluated is the quadratic deviation :

$$e(t0, a) = \int_{ti}^{ti+TEMPO} (x(t) - x^*(t - t0) - a)^2\, dt \qquad (2)$$

The OP calculus cost is low, for x(t) and x*(t) are piecewise linear functions.
The minimization of e (t0,a) is simplified by writing (2) as :
$a^2 * TEMPO - 2 * a * \varepsilon(t0, 0) + e(t0, 0)$, where ε(t0, z) is defined by :

$$\varepsilon(t0, z) = \int_{ti}^{ti+TEMPO} (x(t) - x^*(t - t0) - z)$$

It enables to separate t0 and a, and to get a quadratic equation for a parameterized by t0. All the evaluations are processed at little cost for the formal calculus is developed as far as possible thanks to the piecewise linear functions. Otherwise the SCHWARTZ inequality enables to link the thresholds of ε(t) and e(t0, a) : if we assume that $|\varepsilon(t)| < e$, the threshold of e(t0, a) will be a percentage of $TEMPO * e^2$.
We can still notice that this calculus can be used to study the parallelism of the signals : if the result of the optimal proximity is a constant vector in time, it means that the evolutions are locally parallel. Two evolutions are said to have similar shapes if a translation (t0 ; a) for which the criterion e (t0, a) is reduced below the given threshold can be found. It corresponds to $e(t0 ; a) \rightarrow 0$ qualitatively after minimization by OP.
Two evolutions are said to be qualitatively similar if at least one of the next conditions is verified :
- there is no error in the sense of (1),
- the evolutions have similar shapes in the sense of (2),
every combination of these two conditions is interpreted.
The aim of the comparison algorithm is to detect as precociously as possible complete failures and drift failures. The first ones are generally characterized by significant drops in the evolution of the error functions and the second ones by smoothly increasing error functions. Other causes of divergences such as imprecision inherent to the modelling or disturbances of the measurement system must not entail a defect detection.

3. INTRODUCTION OF A FUZZY ANALYSIS

3.1. Vague facts and fuzzy set theory

For our purpose, we use the numerical values computed by (1) and (2) but they have to be interpreted symbolically to lead to a decision ;
- the simulation is identical with the process evolution, there is no problem ;
- the simulation is very different from the process evolution, a diagnosis has to be started ;
- the simulation does not exactly correspond to the process evolution, a more precise study has to be carried out before a decision is taken.

For instance, the error ε(t) could be 23% whereas the corresponding decision rule could be : "if the error is high, a diagnosis is started". If the attribute "HIGH" is based on a single threshold, the system using this decision rule may react very differently if
ε(t) = THRESHOLD_VALUE + δ or
ε(t) = THRESHOLD_VALUE - δ, regardless of the magnitude of δ.
The difficulty here does not come from the improper choice of the value of the threshold THRESHOLD_VALUE but from the lack of existence of a well defined threshold that might separate error values compatible with "HIGH" from values that are incompatible. "HIGH" is a vague predicate, and should be modelled as such [6]
ZADEH's fuzzy set theory offers a very simple and elegant tool for dealing with vagueness of terms, especially when they clearly refer to one or several numerical scales [7].

Using this approach, a vague condition such as "error ε (t) is high" is represented by means of fuzzy intervals as represented in figure 3 :

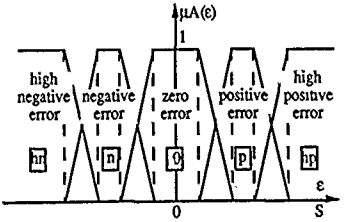


Figure 3 . Fuzzy intervals

The relevance of the terms F: HIGH NEGATIVE (HN), NEGATIVE (N), ZERO (0), POSITIVE (P) and HIGH POSITIVE (HP), with regard to the numerical value of $\dot{\varepsilon}(t)$, is expressed as a compatibility function. It means that for any value $\dot{\varepsilon}(t) \in S$, a degree $\mu_F(\dot{\varepsilon}) \in [0, 1]$ can be obtained that evaluates the relevance of the term F to the situation described by the value $\dot{\varepsilon}$ (or equivalently, the degree of compatibility of $\dot{\varepsilon}$ with F). Note that $\mu_F(\dot{\varepsilon})$ is not a degree of uncertainty. Indeed the situation is completely known ($\dot{\varepsilon}(t) = 23\%$).

3.2. Degrees of relevance

The various degrees of relevance or degrees of compatibility used in our comparison algorithm are now presented. Each calculated or measured value $\dot{\varepsilon}(t)$, or $\dot{\varepsilon}(t0, a)$, ... is characterized by a degree of relevance : $\mu_A(\dot{\varepsilon})$, or $\mu_A(e)$, ..., which associates it with its compatibility with the vague predicate A.

$\mu_A(\dot{\varepsilon})$ is associated with the criterion (1), where A may be the symbolic value HIGH POSITIVE (HP), POSITIVE (P), ZERO (0), NEGATIVE (N) or HIGH NEGATIVE (HN).

Two other degrees are affected to the criterion $\dot{\varepsilon}(t0, a)$, (2). The optimal proximity algorithm provides not only the evaluation of $\dot{\varepsilon}(t0, a)_{min}$, the minimal quadratic deviation between the segmented evolution and the simulated one, but also the translation vector that gives this minimum. The norm of this vector corresponds to the cost of the optimal proximity, the importance of the translation to be carried out is given by (3) : $\frac{1}{\sqrt{2}} \cdot \sqrt{\frac{t0^2}{\text{MARGIN}^2} + \frac{a^2}{\delta^2_{max}}}$ (3),

where MARGIN and δ_{max} correspond to the components of the largest allowed translation. Consequently the minimal quadratic deviation and the corresponding cost are associated with the interest of the optimal proximity calculus.

$\mu_A(e)$ corresponds to the compatibility of $\dot{\varepsilon}(t0, a)$ with the symbolic value A, which may be HIGH (H) or ZERO (0).

$\mu_A(c)$ corresponds to the compatibility of the translation cost with the symbolic value A, which may be HIGH (H) or ZERO (0).

The interest of the optimal proximity can be represented through a symbolic scale that takes into account at one and the same time the minimal quadratic deviation e and the cost c of the operation. The degree μ_A (OP) is a function of $\mu_A(e)$ and $\mu_A(c)$. A may be "BAD" (if the optimal translation has not resulted in similar shapes) or "OK" in other case. Let us set :

- when the error (1) is P or N :

$$\mu_{BAD}(OP) = \mu_H(e) + \mu_H(c) - \mu_H(e) \bullet \mu_H(c)$$

$$\mu_{OK}(OP) = \mu_0(e) \bullet \mu_0(c)$$

- when the error (1) is HP or HN : $\mu_{BAD}(OP) = 1$ and $\mu_{OK}(OP) = 0$.

The decision to start a diagnosis is represented as a symbolic scale that expresses the risk that the situation is incidental whereas it corresponds to a knowledge expressed by terms in the natural language. μ_A (DECISION) is the weight attributed to a decision : A may be DIAGNOSIS or OK. Let us set :

$$\mu_{DIAGNOSIS}(DECISION) = \mu_{OP}(BAD) \bullet \mu_E(P \text{ or } N),$$

with regard to the definition of two similar evolutions, and

$$\mu_{OK}(DECISION) = 1 - \mu_{DIAGNOSIS}(DECISION).$$

3.3. Results

To detect complete failures, the deviation between the simulated evolution and the measured one is generally so large that the optimal proximity algorithm is unable to reduce enough the quadratic deviation in the permissible part of the plane. So the cost c or the minimal quadratic deviation are usually very high.

With regard to the drift failures, the result of the optimal proximity algorithm is generally successful but when the test is repeated, the cost of the calculus is increasing in time (increasing norm of the optimal proximity translation vector).

Other error sources and noises generally correspond to low cost or low $\dot{\varepsilon}(t0, a)_{min}$.

5. DISCUSSION

The comparison or defect detection algorithm is a local treatment where each variable is individually studied as in classical alarm systems. We are now working on an over refinement of the previous step : a pre-diagnosis. When each variable has been individually examined by the comparison task, the pre-diagnosis is started when simulated and real evolutions of at least one variable are considered to be significantly different by the defect detection. The idea is originally that the defect proposed by the comparison is not necessarily a primary defect : it may not correspond to the genuine source of the malfunction. The aim of the pre-diagnosis is to find the root variable from the detection variable. The analysis rests on the causal graph supporting the simulation. It allows the focalization on a subgraph which is suspected to be in fault. This subgraph is a consistent path whose root is a candidate for the origin of the malfunction.

The aim of the modelling by Qualitative Responses is to describe the global behaviour of a process. This model is not based on a sound physical analysis of the various phenomena occurring in the process : it provides an approximation of the behaviour of each variable in the causal graph, whose abstraction level is more adapted to supervision tasks. A specific representation of signals has been created thanks to the notion of event, and the shape of a signal is as important as its values, that is why a specific defect detection algorithm had to be developed to check consistency of the complete system. It is based on fuzzy calculus for various quantities, such as instantaneous or integral errors.

Furthermore, the strength of the algorithm is due to its weak constraints (qualitatively similar evolutions) that do not lead necessarily to defect suspicion even when a classical simulation error is over a given threshold. The choice of thresholds is then less drastic : they may be set without any accurate quantitative information.

However it is clear that the robustness of the algorithm is more important than precociousness for it has to work on line by taking into account all the disturbances of the measurement system and the modelling approximations.

BIBLIOGRAPHY

[1] J.M. PENALVA, L.COUDOUNEAU, L.LEYVAL, J.MONTMAIN
DIAPASON : Un système d'aide à la supervision soumis aux *Journées Internationales sur les Systèmes Experts et leurs Applications*, Avignon, 1991

[2] L.LEYVAL, S FERAY-BEAUMONT, S GENTIL
Event oriented versus interval oriented qualitative simulation
Mathematical and Intelligent Models in System Simulation, IMACS-IFACS Symposium, Bruxelles, Septembre 1991

[3] J.MONTMAIN, L.LEYVAL, S.GENTIL
On line qualitative interpretation of dynamic simulation for diagnostic
Mathematical and Intelligent Models in System Simulation, IMACS-IFACS Symposium, Bruxelles, Septembre 1991

[4] L. COUDOUNEAU
Un générateur de système expert hypothético-déductif
Mémoire de DEA - Laboratoire DELIA - INSA Lyon, Septembre 1989

[5] I. BLOCH-BOULANGER, H. MAITRE, M. MINOUX
Optimal matching of 3-D convex polyhedra
International Journal of Computer Vision, 1989

[6] D. DUBOIS, H. PRADE
Handling uncertainty in expert systems : pitfalls, difficulties, remedies
Séminaire sur la sécurité et les risques dans l'utilisation des systèmes experts, Copenhague, 1988

[7] L.A.ZADEH
The concept of a Linguistic Variable and its Applications to Approximate Reasoning
Information Sciences n° 8, pp. 199-249, 1975

Qualitative Operators for Order of Magnitude Calculus: Robustness and Precision

N. Piera
C.E.A.B-C.S.I.C.
Camí de Santa Bàrbara s.n.
17300-Blanes, Spain.

M. Sánchez
Dept. de Matemàtica Aplicada II
Facultat d'Informàtica
Pau Gargallo 5, 08028-Barcelona, Spain.

L. Travé-Massuyès
L.A.A.S-C.N.R.S.
7, Avenue du Colonel Roche
31077-Toulouse, France.

Abstract. Order of Magnitude Qualitative Algebras are built on a partition of the real line and use Q-operators which allow some Qualitative calculus which is required to be consistent with real numbers and operations. However the tables of the Q-operators depend on the position of the boundary numbers determining the partition. In this paper, a study of all the possible symbolic tables for the Q-sum is performed. Robustness and precision of these tables are analized. Moreover, it is shown that the boundaries can vary up to some limit still keeping the Q-sum table invariant. Finally, a particular attention is payed to the much simpler symmetric partition case, which is useful in many practical applications

I. Introduction

Qualitative models are characterized by a middle position between heuristic models, requiring solely symbolic manipulations, and pure numerical models, using exclusively calculations. As long as engineering problems are considered, one deals with physical quantities. The qualitative algebraic structure should thus remains consistent with real numbers and operations. The specific class of *Order of Magnitude Qualitative Algebras*, which are built from partitions of the real line, have been defined this way ([4],[7]). This class includes the usual Signs Algebra based on the roughest partition of $\Re$. $+, 0, -$.

However, the tables of Q-operators are not independent of the specific partition of the real line we are working with. This means that if the position of the numbers determining the partition are changed, the table for the Q-operator might be different. In this paper, a whole study is performed for the Q-sum However, it is important to mention that the procedure remains valid for any other Q-operator, the Q-product in particular. A study of *robustness* tables is made, and it is shown that a radius around the boundary numbers can be found so that the variations of the boundaries up to this radius still preserve the same symbolic table.

II. Preliminaries: Q-pairs defined on the real line

Definition II.1 Consider a nonempty set S (*Univers of description*) and an order relation $\leq$ defined on S. *Qualitative equality* $\approx$ (Q-equality) is defined as the following binary relation induced by $\leq$ on S: $a \approx b$ if there exists $x \in S$ such that $x \leq a$ and $x \leq b$.

A *Qualitative pair* $(S, \approx)$ is defined as a set S and a Q-equality defined on S.

In practice, the problem is generally to represent physical quantities which are known with poor precision so that real numbers cannot be used Qualitative models provide a solution as long as they remain consistent whith real numbers and real operators. One way to do so is to define the Q-pairs from a partition of the real line. This paper considers the Q-pair obtained from partitioning the real line $\Re$ into seven classes with associated labels: negative large (NL), negative medium (NM), negative small (NS), zero (0), positive small (PS), positive medium (PM) and positive large (PL) (see Fig. 1). The set $S_1 = \{NL, NM, NS, 0, PS, PM, PL\}$, ordered by $NL < NM < NS < 0 < PS < PM < PL$, constitutes the highest level of specification of our Q-model, The lowest one is total undetermination (?). Between the highest level of specification and the lowest one given by ?, the actual partitioning induces four ordered levels of specification. Interpreting the labels as intervals, the four levels correspond to the union of two, three, four, and five adjacent intervals. Labels within these levels are denoted $[\alpha, \beta]$, where $\alpha, \beta \in S_1 - \{0\}$ and $\alpha < \beta$. $[\alpha, \beta]$ represents the interval obtained from the union of the intervals associated with α and β and the ones in between Our universe of description is thus $S = S_1 \cup \{[NL, NM], [NM, NS], \ldots, [NL, PM], [NM, PL], ?\}$ and the order relation $\leq$ considered in S is the inclusion.

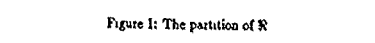


Figure 1: The partition of $\Re$

Definition II.2 Let $\otimes$ be a binary operation on S and let $\times$ be a binary operation on $\Re$. $\otimes$ is said to be *consistent with* $\times$ if for any $\alpha, \beta \in S$, $\alpha \otimes \beta$ is the minimum set of S (with respect to the inclusion) that contains the interior of $\alpha \times \beta$.

III. The Qualitative sum

From now on, focus is put on the particular case of the Q-sum $\oplus$. However it is important to mention that a similar study can be done for any Q-operator (the Q-product for instance). The important concept of consistency of Q-operators does not go without problems. In particular, the qualitative result may differ when the boundaries of the intervals are changed For example, if 10 is the upper boundary of PS, then $PS \oplus PS = [PS, PM]$ if the upper boundary of PM is greater than or equal to 20, and if it is less than 20 then $PS \oplus PS = +$. Altough there are some invariant results, like for example $NL \oplus NL = NL$, the symbolic table for the operation $\oplus$ depends on the real numbers a, b, c, d defining the partition S_1 All possible symbolic tables for the Q-sum $\oplus$ are included in Table 1 (half of it only has been filled because $\oplus$ is strictly commutative).

$\oplus$	NL	NM	NS	$+$	PS	PM	PL
NL	NL	NL	NL	NL	$[NL, NM]$ $-$ $[NL, PS]$	$[NL, PS]$ $[NL, PM]$	?
NM		$[NL, NM]$ NL	$[NL, NM]$	NM	$[NM, NS]$ $[NM, PS]$	$[NM, NS]$ $[NM, PM]$ NS $[NS, PS]$ $[NS, PM]$ PS $[PS, PM]$	$[NM, PL]$ $[NS, PL]$ $[PS, PL]$ $[PM, PL]$
NS			$[NM, NS]$	NS	$[NS, PS]$	$[NS, PM]$ PS $[PS, PM]$	$[NS, PL]$ $+$ $[PM, PL]$
$+$				$+$	PS	PM	PL
PS				$+$ $[PS, PM]$	$[PM, PL]$	PL	
PM					PL	$[PS, PL]$	
PL							PL

Table 1: All the Q-sum tables

From the qualitative point of view, two symbolic Q-sum tables are said to be Q-equals if every couple of corresponding squares are Q-equals. Taking into account this definition, it is easy to see that any $\oplus$ table is Q-equal at least to one of the two tables included in Table 2 . So, it can be concluded that there only exist two tables qualitetively differents

Nevertheless, not all the combinations in Table 1 are possible, for example $PS \oplus PS = +$ is incompatible with $PM \oplus PM = [PM, PL]$, since $PS \oplus PS = +$ if and only if $d < 2c$ and $PM \oplus PM = [PM, PL]$ if and only if $d > 2c$

$\oplus$	NL	NM	NS	$+$	PS	PM	PL
NL	NL	NL	NL	NL	$[NL, NM]$	NS	?
NM		NL	$[NL, NM]$	NM	$[NM\ NS]$		$[PM, PL]$
NS			$[NM, NS]$	NS	$[NS, PS]$	$[PS, PM]$	$[PS, PL]$
$+$				$+$	PS	PM	PL
PS					$[PS, PM]$	$[PS, PL]$	PL
PM						PL	PL
PL							PL

Table 2: Qualitetively different $\oplus$-tables

In the following, our first problem is to discern how many tables are possible and to determine them Given $(a, b, c, d) \in R^4$ with $a < b < 0 < c < d$, then the symbolic table of $\oplus$ is completely determined. Whence if we consider $D = \{(a, b, c, d) \in R^4, a < b < 0 < c < d\}$ we can establish the following equivalence relation:

Definition III.1 Two elements of D are said to be $\oplus$-*equivalents* if they induce the same table of $\oplus$.

IV. The conditions related to Q-sum table

Taking into account Fig. 1, the results in non fixed squares of Table 1 depend on the relations of a, b, c, and d For instance, the result of $NL \oplus PS$ is conditioned in the following way.

$$NL \oplus PS = \begin{cases} [NL, NM], & \text{if } c \leq b - a, \\ -, & \text{if } b - a < c \leq -a, \\ [NL, PS], & \text{if } -a < c. \end{cases}$$

Proceeding in a similar way for the eleven undetermined squares, the conditions determining the table can be summarized in an algorithmic way as follows:

1. Compare b with $\frac{a}{2}$.

2. Compare c with $b-a$, $-b$ and $-a$. Depending on the result of applying the first rule, the situation of these three numbers $b-a$, $-b$ and $-a$ is different. We have the three following possibilities in the positive half real line:

   i) If $b < \frac{c}{2}$, then $0 < b-a < -b < -a$;

   ii) If $b = \frac{c}{2}$, then $0 < b-a = -b < -a$;

   iii) If $b > \frac{c}{2}$, then $0 < -b < b-a < -a$.

   Then, depending on the situation the possibilities for c are different

3. Compare d with those of these numbers $b-a$, $-b$ and $-a$ which are greater than c and with $c-b$, $c-a$ and $2c$.

The arborescence of the above algorithm provides all the possible Q-sum tables. The number of possible different tables is 201.

V. The $\oplus$-equivalence classes. Robustness and Precision.

From the outlined conditions it is evident that any $\oplus$-equivalence class corresponds to a connex domain of D. So, given a point p in D, there always exists a convenient direction of move from p such that the associated table remains invariant. On the other hand, there exist $\oplus$-equivalence classes with hypervolum 0, and another kind of classes with infinite hypervolum. Of course, if the $\oplus$-equivalence class of a point $p \in D$ has hypervolum 0, i.e. it is a strengh line, only one direction of move from p preserves the Q-sum table. If the $\oplus$-equivalence class of p has infinite hypervolum, either a sector of directions of move from p preserve the table if p is on the frontier of the class, or all directions if p is an interior point of its class.

V.1 Robustness. Given two tables T_i and T_j, the question is when is T_i more robust than T_j? That is, if C_i and C_j denote their associated $\oplus$-equivalence classes, when is the domain C_i "greater" than the domain C_j? In order to compare the "size" of these domains, we compare first the possible results of the sum $\alpha \oplus \beta$ for any pair $(\alpha,\beta) \in S_3^4$, that is for every square of the table. The robustness of the whole table is a consequence of the stability of the result of the sum within every square. Of course, fixed squares are not considered. Since some conditions give rise to infinite hypervolumes, the comparison is performed by using a geometrical criterium.

Given $(\alpha,\beta) \in S_3^4$, let x and y be two possible results of $\alpha \oplus \beta$, then x is *more stable than* y if the domain $D_x \subset D$ corresponding to the result x contains the image of the domain $D_y \subset D$ by a rigid movement of R^4. A table is said to be *more stable than* another, if all the results of the pairs $\alpha \oplus \beta$ corresponding to the first are more stable than those that corresponding to the second one.

The relation *to be more stable than* between tables is an order relation, but is not a total order. Table 3 shows the most stable Q-sum table and $C_M = \{(a,b,c,d) \in D;\ \frac{c}{2} < b, c > -a, d > 2c\}$ is it associated $\oplus$-equivalence class.

$\oplus$	NL	NM	NS	0	PS	PM	PL
NL	NL	NL	NL	NL	[NL, PS]	[NL, PM]	?
NM		[NL, NM]	[NL, NM]	NM	[NM, PS]	[PS, PM]	[PM, PL]
NS			[NM, NS]	NS	[NS, PS]	[PS, PM]	[PM, PL]
0				0	PS	PM	PL
PS				PS	[PS, PS]	[PM, PL]	PL
PM				PM		[PM, PL]	PL
PL				PL			PL

Table 3: The most stable Table

V.2 Precision. Another aspect to consider is the *precision* of the Q-sum tables. This concept is related to the fact that the result of $\alpha \oplus \beta$ for $(\alpha,\beta) \in S_3^4$ may be an element of S_1 or not. In the first case, it has the maximal precision allowed by our algebra. Hence, the more information a table provides, the more precise it can be said. As robustness, the precision of the tables is studied "square by square" by comparing the results of the sum $\alpha \oplus \beta$ for any of the pairs $(\alpha,\beta) \in S_3^4$ which do not correspond to fixed squares. The precision of the tables is a consequence of precision of each of the possible results in the squares. However a measure for the degree of precision of the result $z \in S$ of $\alpha \oplus \beta$ can be provided in the following way. a) if $z \in S_1$, then z has *degree of precision* 1; b) otherwise, the *degree of precision* of z is $\overline{Card(B_{\alpha\oplus\beta} - \{0\})}$, where $B_{\alpha\oplus\beta} = \{z \in S_1; z \approx \alpha \oplus \beta\}$.

From these *marginal* degrees, the *degree of precision of the whole table* is obtained by taking the minimum of the degrees of precision of all the no fixed squares of the table. A table is said to be *more precise than* another, if the degree of precision of the first is greater than the degree of precision of the second. This relation is a pre-order that has neither a maximum nor a minimum.

VI. Sensitivity analysis

This section deals with the problem of finding how the boundaries of the intervals of the partition of $\Re$ can freely move, in a continuous way, keeping the symbolic Q-sum table invariant. From the topological point of view, two kinds of $\oplus$-equivalence classes must be distinguished: the open classes and the ones that have the property that all their points are frontier points. First, the study is performed for the Q-sum tables with open associated $\oplus$-equivalence class.

Let $p = (a_0,b_0,c_0,d_0)$ be a point of D such that its $\oplus$-equivalence class C_p is an open set in R^4. Moving the four coordinates of p in a continuous way and still remaining inside of C_p, can be interpreted as the research of a common radius r for this four coordinates, such that every coordinate can vary inside its corresponding interval (See Figure 2). This is equivalent to

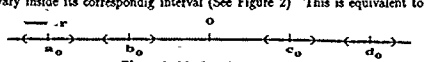


Figure 2: Moving the boundaries

searching an hypercube with center p and side $2r$ totally contained in the class C_p, thas is:

$$S_r(p) =]a_0 - r, a_0 + r[\times]b_0 - r, b_0 + r[\times]c_0 - r, c_0 + r[\times]d_0 - r, d_0 + r[.$$

Moreover, it is possible to find the maximum r such that $S_r(p) \subset C_p$. To this end, it is sufficient to impose that the eigth boundaries of the intervals of variation in Figure 2 satisfy the results of the conditions corresponding to C_p.

Finally, we analyze the case in which C_p is a set such that all their points are frontier points. Then at least one of the conditions that determine C_p is an equality. Searching for a radius r for free movement of a, b, c, d, implies that all the coordinates that take part in any equality condition must remain fixed, and the radius can only be found for the other coordinates.

References

[1] *Artificial Intelligence.* Special volume on Qualitative Physics, Vol 24, 1984.

[2] Dague,P., Raiman, O., and Devés, P. (1987) Troubleshooting: when modelling is a trouble. *Sixth National Conference on Artificial Intelligence*, Seattle, U.S.A.

[3] Dormoy, J.L. (1988) Controlling qualitative resolution. *Second Qualitative Physics Workshop*, Paris.

[4] Piera, N. and Travé-Massuyès, L. (1989) About qualitative equality, axioms and properties. *9th International Workshop on Expert Systems and their Applications*, Avignon, France.

[5] Raiman, O. (1986) Order of magnitude reasoning. *Fifth National Conference on Artificial Intelligence (AAAI-1986)*, Philadelphia, USA.

[6] Simmons, R. (1988) Commonsense arithmetic reasoning. *AI in Engeneering*, 3(3), 156-169.

[7] Travé-Massuyès, L. and Piera, N. (1989). The orders of magnitude models as qualitative algebras. *11th IJCAI*, Detroit, USA.

[8] *Revue d'Intelligence Artificielle* Special volume on Qualitative Reasoning, Vol 3, n. 4, 1989.

Appendix: The symmetric case

Numerous practical problems can be modelled by using a symmetric partition which provides enormous advantages in terms of complexity. This case corresponds to the boundaries of the intervals satisfying $-a = d = \beta$ and $-b = c = \alpha$. There are only three different tables for $\oplus$, instead of 201 in the general case[1], and they are qualitatively equals. These tables are in correspondance with the three domains in the plane $O\alpha\beta$ represented in Figure 3. Since $\theta_2 > \theta_1$, the first two domains are injectable in the third, and therefore the table that corresponds to the case $\beta > 2\alpha$ is the most robust one.

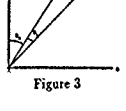


Figure 3

Fundamental Limitations of Qualitative Simulation

Andrej Makarovič Nicolaas J.I. Mars

University of Twente, Department of Computer Science,
P.O.Box 217, 7500AE Enschede, the Netherlands

The purpose of our research is to assess the fundamental limitations of qualitative simulation. In the past decade, several research groups [1], [2], [3] developed simulation techniques based on a common principle which we call Qualitative Simulation. Based on their empirical results, we decided to investigate whether gradual further developments of these techniques could ever lead to acceptable results.

We identified three combinations of fundamental design decisions in Qualitative Simulation, each of which is characteristic for the approach, is hard to revise, and has important negative consequences. In our opinion, those consequences are hardly ever acceptable in practice. The least radical revisions required to overcome these limitations amount to a complete redesign.

In the following sections, these (combinations of) design decisions and their consequences will be disussed briefly. For more information on this topic we refer to the first part of our Ph.D. Thesis [4].

1 Only Signs of Derivatives

A state of a system with n state variables can be visualized by a point in an n-dimensional state space. The solution $\vec{x}(t)$ of a set of the differential equations $\frac{d\vec{x}}{dt} = \vec{F}(\vec{x})$ can be visualized by a so-called trajectory.

The initial set of ordinary differential equations is satisfied iff both the *signs* and the *absolute values* of the left and right terms are equal, i.e., $\widetilde{Sign}(\frac{d\vec{x}}{dt}) = \widetilde{Sign}(\vec{F}(\vec{x}))$ and $\widetilde{Abs}(\frac{d\vec{x}}{dt}) = \widetilde{Abs}(\vec{F}(\vec{x}))$. In Qualitative Simulation, the equality of the absolute values is not verified, hence the direction of each piece of a trajectory can be known only up to an n-dimensional generalisation of a quadrant accurately. A quickly diverging bundle of "possible" trajecories is thus predicted. The predicted trajectories seldom have interesting properties in common, hence little is actually predicted about the exact solutions.

2 State Space on Ordinal Scale & Markov Sequence of Short Term Predictions

In Qualitative Simulation, long term predictions are formed by independently chaining short term predictions, just like in the traditional numeric aproaches. The state variables are represented on an ordinal scale. By definition, on an ordinal scale only the order of the values is represented, the lengths of the intermediate intervals are unknown.

A problem with this combination of design decisions is that the lenghts of the short term displacements can not be represented as an intermediate result. However, in general, this information is required for determining the signs of the composed long term displacements.

$$\begin{aligned} Sign(x_i(t_2) - x_i(t_0)) &= Sign\left(\frac{x_i(t_2) - x_i(t_1)}{x_i(t_1) - x_i(t_0)} + \right) \\ &= Function\left(\begin{array}{l} Sign(x_i(t_2) - x_i(t_1)), \\ Sign(x_i(t_1) - x_i(t_0)), \\ Abs(x_i(t_2) - x_i(t_1)), \\ Abs(x_i(t_1) - x_i(t_0)) \end{array}\right) \\ &\neq Function\left(\begin{array}{l} Sign(x_i(t_2) - x_i(t_1)), \\ Sign(x_i(t_1) - x_i(t_0)) \end{array}\right) \end{aligned} \quad (1)$$

A consequence is that ordinal relationships of a *state* variable as a function of time can only be determined within the same period of monotonic behavior. The ordinal relationships between non-adjacent local extremes, like two adjacent local maximums, thus can not be determined. Another consequence is that a revision of the first combination of design decisions (only signs of derivatives) can not have any favorable effect unless this second combination of design decisions is revised.

3 Prediction Step-Size & Orthogonal Quantification of State Space

In the traditional approaches, the state space is quantified by a grid, the resolution of which is determined by the number of digits used. In Qualitative Simulation, the state space is similarly quantified by an orthogonal grid into so-called qualitative states. For example, in the 3-dimensional state space, a qualitative state is a cube, bounded by 26 qualitative boundaries (6 sides, 12 edges and 8 corners).

The prediction step-size is the length of a displacement of a short term prediction. $|\vec{x}(t_i) - \vec{x}(t_{i-1})|$ In Qualitative Simulation, the prediction step size has been chosen so small that short term predictions correspond with transitions between *adjacent* qualitative states.

This third combination of design decisions has three consequences:

A revision of the first combination of design decisions (only signs of derivatives) can not have any favorable effect unless this third combination of design decisions is revised. A state transition between two adjacent qualitative states is possible iff a trajectory exists that successively visits those qualitative states This is determined by the direction of the derivative vector at the intermediate qualitative boundary Because of the orthogonal quantification of the state space, that direction must be known only up to an n-dimensional generalization of a quadrant accurately, only the signs of the derivatives can thus be used effectively.

The predictions are drowning in rounding-off errors Rounding off errors are of the same magnitude as the diameters of the qual

itative states. Because the prediction step-size has been chosen equal to the diameter of the diameters of the qualitative states, the short term displacements are drowning in rounding-off errors.

The average (averaged over initial qualitative states) number of possible qualitative successor states per qualitative state is $BF \approx 2^n - 1$, where n is the number of state variables. If we assume that the state space has not been quantified along trajectories, then this number is independent of the actual vector field (i.e., the set of differential equations). The Branching Factor BF is defined as the quotient of the total number of Possible State Transitions $\sharp PST$, and the total number of Qualitative States $\sharp QS$.

$$Branching\ Factor = \frac{\sharp PST}{\sharp QS} \approx \frac{\sharp PST}{\sharp QB} \cdot \frac{\sharp QB}{\sharp QS} \qquad (2)$$

In favor of Qualitative Simulation, let us assume that each qualitative boundary is intersected by trajectories in only one direction, hence $\frac{\sharp PST}{\sharp QB} \approx 1$. The number of qualitative boundaries per qualitative state depends only upon the dimension of the state space n: $\frac{\sharp QB}{\sharp QS} \approx 2^n - 1$. The average number of predicted qualitative behaviors is thus

$$\sharp Predicted Behaviors \approx BF^s \approx (2^n - 1)^s \qquad (3)$$

where s is the numer of prediction steps.

4 Misconceptions

The article of Fouché & Mélin [5] seems to contradict with our results twice: with respect to determining behavior over periods of non-monotonic behavior, and with respect to the practical usefullness of Qualitative Simulation.

4.1 The first misconception

Fouché & Mélin have shown that Qualitative Simulation in combination with an appropriate model can determine the decreasing amplitude of the the oscillations of a non-linear damped oscillator.

We claim that as a consequence of the second combination of fundamental design decisions, ordinal relationships of a *state* variable as a function of time can only be determined within the same period of monotonic behavior. Consequently, if "position" and "velocity" are state variables of that oscillator, then their decreasing amplitude can not be determined by Qualitative Simulation.

A state variable is a variable only the derivative of which is determined by the (ordinary) differential equations $\frac{d}{dt} StateVar = F(StateVars, InputVars)$; no instantaneous relationships between state variables can exist. Instantaneous relationships determine the values of non-state variables $Non - StateVar = G(StateVars, InputVars)$.

Ordinal relationships of *non-state* variables as a function of time can be specified, even over periods of non-monotonic behavior. Let us assume that there is only one state variable called "time". That time is specified to increase monotonically $\frac{d}{dt} time = Constant > 0$. The function $G(time)$ can specify any behavior.

Although the proposal of Fouché & Mélin is more sophisticated, it is nevertheless based on the same principle of specifying a monotonically decreasing state variable called "energy". The variables "position" and "velocity" are instantaneously related with the variable "energy", and thus can not be state variables any more. Our claim thus does not contradict with the experimental result of Fouché & Mélin

4.2 The second misconception

Fouche & Mélin claim that their example shows that Qualitative Simulation can solve a useful non-trivial problem. We agree with their statement that the solution is useful and non trivial. However, we disagree with the statement that the solution has been found by Qualitative Simulation.

The major contribution for finding that solution consists of the preparatory manual analysis of the system. We encourage the authors to continue automating that part. However, we think that the subsequent use of Qualitative Simulation is just a detour.

5 Conclusion

We identified three combinations of fundamental design decisions in Qualitative Simulation, each of which is characteristic for the approach, is hard to revise, and has important negative consequences. In our opinion, those consequences are hardly ever acceptable in practice. The least radical revisions required to overcome these limitations amount to a complete redesign.

Qualitative Simulation is practically useless and irreparable, and should therefore be abandoned.

References

[1] B. Kuipers, "Qualitative Simulation," *Artificial Intelligence* 29 (1986), 289-338.

[2] J. De Kleer & J.S. Brown, "A Qualitative Physics Based on Confluences," *Artificial Intelligence* 24 (1984), 7-83.

[3] K.D. Forbus, "Qualitative Process Theory," *Artificial Intelligence* 24 (1984), 85-168.

[4] A. Makarović, "Parsimony in Model-Based Reasoning," Ph D. Thesis, University of Twente, 1991.

[5] A.C.P. Fouché & C. Mélin, "Recent Improvements In Qualitative Simulation," *Proceedings of the 13th IMACS World Congress*, Dublin (1991).

Recent Improvements In Qualitative Simulation†

Anne Charles, Pierre Fouché & Christian Mélin
Université de Technologie de Compiègne, Département Génie Informatique, U.R.A. C.N.R.S. 817
Centre de Recherches de Royallieu, BP 649, 60206 Compiègne cédex, FRANCE.

Abstract: This paper presents some problems encountered during qualitative simulation of dynamic systems, reviews existing solutions and illustrates them on a non-trivial example.

Introduction

Qualitative simulation is an approach for deriving behavioral information about physical systems that are usually modeled by systems of ordinary differential equations. Several researchers (for instance Caloud [1], Forbus [3], de Kleer & Brown [2], Kuipers [6], Shen & Leitch [13], Sacks [12], Williams [14]) have developped qualitative simulators. From our point of view, the most domain-independent, mathematically sophisticated algorithm is Kuipers' QSIM [6], [9]. However some objections have been formulated against the foundations of qualitative simulation [11] and it has been claimed that it would always be impossible to simulate any non-trivial system. The goal of this paper is to analyze the main problems encountered in qualitative simulation, to briefly review the techniques that address these problems and to show on a non-trivial example that QSIM can derive useful conclusions about it.

Qualitative Simulation: Problems, Causes and Solutions

Very often first runs of qualitative simulation produce a lot of behaviors, from which it is difficult to extract behavioral features of a system. Behaviors may proliferate either because they are not represented at the appropriate level of description (in which case they differ very slightly) or because spurious behaviors are predicted (i.e. behaviors that do not correspond to any solution of any system of differential equations consistent with the qualitative model). In practice, when one analyses an intractable tree of behaviors, one can observe two phenomena, referred to as chatter and occurrence branching, that are the major symptoms of the two problems:

- A variable may exhibit chatter if its derivative is unconstrained. Basically, if at some time point a variable transitions to a critical point (that is, its derivative becomes zero) then its qualitative value in the next open interval of time is determined by its second derivative. If no information is provided about this second derivative then simulation will branch on each possible future
- Occurrence branching happens when the temporal ordering of two or more events (for instance, a variable crossing a landmark or reaching a critical point) cannot be determined. Qualitative simulation then branches on all the possible orderings.

Causes of behavior proliferation are found in the main assumptions which qualitative simulation is based on, and can be classified as follows:

1. Representation of a variable's values:
   1. Discretization: the domain of a variable is partitioned into "landmark" points and intervals between them,
   2. Ordinal scale: only the ordering of landmarks is represented.
2. Representation of a variable's derivatives: in basic algorithms, only the sign of the first derivative is represented, and higher order derivatives are ignored.
3. Local reasoning: successors of a given state are determined using only that state, without considering any preceding state,
4. Model specification: simulation results are very sensitive to the choice of variables and landmarks, as well as constraints and corresponding values.

In the following section we explain how symptoms and causes are related, and present existing solutions.

Chatter

As mentioned above (cause #2), only the sign of the first derivative is explicitly represented. As long as no information about the second derivative of a variable is provided, simulation will branch whenever a variable reaches a critical point.

Possible solutions are:

- add a new variable to represent the derivative of a chattering variable, and constrain it. This would only translate the problem, because the derivative of the newly introduced variable would exhibit chatter as well;
- find an expression for the second derivative of a variable, and use it when its first derivative becomes zero* [7];
- ignore the direction of change of a variable so critical points can no longer be detected* [7];
- partition the envisionment graph into equivalence classes to gather similar states* [4].

Occurrence Branching

Indetermination in temporal ordering of events may be caused by limited knowledge about landmarks (cause #1.2 in conjunction with cause #2), an inappropriate model specification (cause #4) or sometimes by the local nature of reasoning (cause #3). Some solutions exist:

- disable the possibility to create new landmarks during simulation for some variables*;
- remove some landmarks from quantity spaces so events defined by a variable reaching or crossing that landmark can no longer happen;
- ignore the direction of change of a variable so events defined by critical points can no longer happen* [7];
- more specify constraints by adding corresponding values;
- analyze similar behaviors and group those which differ only by temporal orderings of events* [4];
- add partial quantitative information* [8].

Other symptoms

Behaviors may also proliferate for reasons that are not easy to understand and which are not associated with particular symptoms. This is mainly due to the local nature of reasoning (cause #3) in conjunction with the discretization of a variable's domain (cause #1.1; it happens that the validity of a transition depends on the behavior in which it takes place), or an inappropriate model specification (cause #4) Possible remedies include:

- disable the possibility to create new landmarks during simulation for some variables*;
- reason in the phase space representation and eliminate behaviors that correspond to self-intersecting trajectories in the phase plane of two independent variables (for second-order systems)* [10].
- reason in terms of energy, decompose processes into conservative and non-conservative ones, and check if the law of conservation of energy is satisfied* [5].
- reconsider the model, add new variables and/or redundant constraints,
- add partial quantitative information* [8].

Using Qsim in Practice

In this section we will consider an example of a second-order, non-linear system for which analytic solutions are not known, but for which one can derive useful information about its solutions, using qualitative simulation. Our goal is to show incrementally how the techniques presented above can be used to come up with a tractable tree of behaviors. The system (Σ) is $\frac{dx}{dt} = -x^3 - y$ and $\frac{dy}{dt} = x - x^2 y$

Eliminating y in (Σ) yields the following equation:

$$\frac{d^2x}{dt^2} = -4x^2\frac{dx}{dt} - (x^4 + 1)x.$$

As $-4x^2 \leq 0$ for all x and $x^4 + 1$ depends only on x, one can make an analogy with a mechanical system, x representing the position of an object, $\frac{dx}{dt}$ its velocity and $\frac{d^2x}{dt^2}$ its acceleration $-4x^2\frac{dx}{dt}$ represents a friction force and $-(x^4 + 1)x$ a conservative

force. Such a system is dissipative and reaches asymptotically an equilibrium state. We will show that this intuitive description of (Σ)'s solutions is confirmed by qualitative simulation performed by QSIM.

We begin with building a qualitative model M_1 of (Σ). The variables are x, y, dx/dt, dy/dt, x2, x3, -y and x2y, and the constraints are d/dt (x dx/dt), d/dt (y dy/dt), minus (y -y), mult (x x x2), mult (x2 x3), mult (x2 y x2y), add (dx/dt x3 -y) and add (dy/dt x2y x).

The way we approached the problem was the following: we simulated the model, analyzed why we came up with too many behaviors, chose an appropriate technique to solve the problem and then reran a simulation. What follows sums up what we obtained:

Simulation results: 18 behaviors at time t_2 (see figure a). Analysis: dx/dt chatters. Solution: ignore dx/dt's direction of change.

Simulation results: 7 behaviors at time t_3. Analysis: x2y chatters. Solution: ignore x2y's direction of change.

Simulation results: 12 behaviors at time t_3. Analysis: dy/dt chatters. Solution: ignore dy/dt's direction of change.

Simulation results: 3 behaviors at time t_2. Analysis: 1 spurious behavior in which $y'(t_2) = 0$ and $x(t_2) \neq 0$ (impossible!). Solution: Modify the model.

Pragmatically with QSIM, it is better to qualitatively translate a factorized equation than a developped one, because two equations algebraically equivalent, but expressed differently, may have non-equivalent qualitative translations. In our case, $\frac{dy}{dt} = x - x^2y$ is equivalent to $\frac{dy}{dt} = x(1 - xy)$, but the set of constraints {mult (x x x2), mult (x2 y x2y), add (dy/dt x2y x)} is not equivalent to the set of constraints {mult (x y xy), add (xy 1-xy 1), mult (x 1-xy dy/dt)}. Consequently we changed the model M_1 to a model M_2 whose variables are x, y, dx/dt, dy/dt, x2, x3, -y, xy and 1-xy, and the constraints are d/dt (x dx/dt), d/dt (y dy/dt), minus (y -y), mult (x x x2), mult (x x2 x3), mult (x y xy), add (xy 1-xy 1), mult (x 1-xy dy/dt) and add (dx/dt x3 -y). We obtained the results summarized below:

Simulation results: 4 behaviors at time t_3. Analysis: xy and 1-xy chatter. Solution: ignore xy's and 1-xy's directions of change.

Simulation results: 4 behaviors at time t_3. Analysis: x2 starts from $x2(t_0)$, decreases and reaches 0, then reaches a new maximum, which can be greater than, equal to or lower than $x2(t_0)$. Uninteresting distinction. Solution: Disable landmark creation for x2.

Simulation results: 9 behaviors at time t_6. Analysis: x oscillates, but QSIM is unable to determine the nature of x's oscillations. Solution: Derive an energy constraint for x: decompose d2x/dt2 into conservative and non-conservative terms.

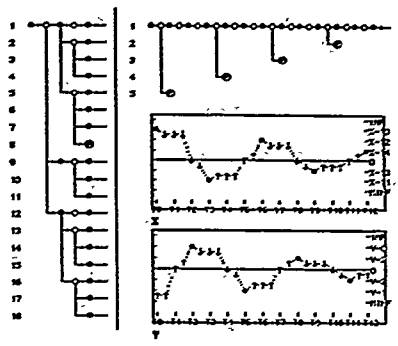
Simulation results: 5 behaviors at time t_6. Analysis: Unable to determine the nature of x3's oscillations: increasing, decreasing or steady. x3 should have a behavior similar to x. But constraint propagation through the means of corresponding values does not happen: no new landmarks are created for x2, thus no new corresponding values for mult (x2 x x3) are created, and x3 is not constrained enough. Solution: Disable landmark creation for x3.

Simulation results: 6 behaviors at time t_8. Analysis. y oscillates, but QSIM is unable to determine the nature of y's oscillations. Solution: Use the phase plane (x y) and eliminate behaviors that self-intersect.

Simulation results: The tree grows linearly: 4 behaviors at time t_8, 8 at t_{20}. Analysis: all the behaviors are possible. x and y exhibit decreasing oscillations. The system can reach an equilibrium state (x = 0, y = 0) after an arbitrary number of oscillations (see figure b).

Conclusion

In this paper, we have shown that QSIM can draw useful conclusions about a system that were not obvious to derive with analytical methods. When the simulation steps we described above will be sufficiently automated, we think that QSIM will be an even more efficient system analysis tool.



(a) (b)

(a) Initial behavior tree
(b) Final behavior tree and x's and y's behavior

References

[1] P. Caloud, "Towards continuous process supervision", in *Proceedings of IJCAI-87*, 1987, pp. 1086-1089.

[2] J. De Kleer, and J. Brown, "A Qualitative Physics based on Confluences", *Artificial Intelligence Journal*, vol. 24, pp. 7-83, 1984.

[3] K.D. Forbus, "Qualitative Process Theory", *Artificial Intelligence Journal*, vol. 24, pp. 85-168, 1984.

[4] P. Fouché, and B.J. Kuipers, "Abstracting Irrelevant Distinctions in Qualitative Simulation", in *Proceedings of the Fifth International Workshop on Qualitative Reasoning about Physical Systems, Austin, TX*, May 1991.

[5] P. Fouché, and B.J. Kuipers, "Introducing Energy into Qualitative Simulation", in *Proceedings of the First European Workshop on Qualitative Physics, Genova, Italy*, January 1991.

[6] B.J. Kuipers, "Qualitative Simulation", *Artificial Intelligence Journal*, vol. 29, pp. 289-388, 1986.

[7] B.J. Kuipers, and C. Chiu, "Taming Intractible Branching in Qualitative Simulation", in *Proceedings of IJCAI-87, Milan, Italy*, 1987, pp. 1079-1085.

[8] B.J. Kuipers, and D. Berleant, "Using Incomplete Quantitative Knowledge in Qualitative Reasoning", in *Proceedings of AAAI-88, Saint Paul, MI*, 1988, pp. 324-329.

[9] B.J. Kuipers, "Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge", *Automatica*, vol. 25, pp. 571-585, 1989.

[10] W.W. Lee, and B.J. Kuipers, "Non-Intersection of Trajectories in Qualitative Phase Space: A Global Constraint for Qualitative Simulation", in *Proceedings of AAAI-88, Saint Paul, MI*, 1988, pp. 286-290.

[11] A. Makarovic, and N. Mars, "Fundamental Limitations of Qualitative Simulation", in *Proceedings of the 13th IMACS World Congress*, 1991.

[12] E. Sacks, "Piecewise Linear Reasoning", in *Proceedings of AAAI-87, Seattle, WA*, 1987, pp. 655-659.

[13] Q. Shen, and R. Leitch, "Integrating Common-Sense and Qualitative Simulation by the Use of Fuzzy Sets", In *Recent Advances in Qualitative Physics*. MIT Press, P. Struss, and B. Faltings, Eds., 1991.

[14] B C. Williams "Doing Time. Putting Qualitative Reasoning on Firmer Ground', in *Proceedings of AAAI-86, Philadelphia, PA*, August 1986, pp. 105-112.

† This work has taken place at the C.N.R.S. U.R.A 817 "Heudiasyc", Université de Technologie de Compiègne. It was supported in part by grants from Atochem and Rhône-Poulenc. It benefited from discussions with Prof. Jean Paul Barthes from the Université de Compiègne and Prof. Benjamin Kuipers from the University of Texas at Austin.

* This technique has been implemented in QSIM and is available in the current released version.

RULE BASED SCHEDULING IN A FIBRE PLANT

Joachim Jachemich
I+K / Software
HOECHST AG
Postfach 80 03 20
D-6230 Frankfurt/Main 80

Abstract : During the last three years an example application of a knowledge based scheduling system has been developed for use in a HOECHST AG fibre plant. The design of the system follows the assumption that plant-level scheduling problems can successfully be modelled using a rule-based approach. In contrast to a mathematical solution, the schedule generated is suboptimal with regard to some (hypothetical) cost function, nevertheless the user acceptance of the system is much more promising than with the previously used "linear programming"-style solution.

I. INTRODUCTION

Starting in the late 60's, the fibre plant under discussion was supposed to use a linear-programming-based system for its scheduling problems. Although the program worked and proposed an "optimum" solution, it turned out to be unsatisfactory with regard to several crucial points :

- Although the initial scheduling proposal was accepted and believed to be optimal, it was almost totally incomprehensible.
- The plant manager was not able to influence the programs behaviour other than through some very rough parameters.
- Inevitable "manual" changes to the proposed schedule usually led to more or less drastically worse solutions.
- The system had a bad response time (several hours).
- The quality of the schedule computed was questionable, since the plant manager usually had several (conflicting) goal functions in mind — there was no way of introducing conflicting goals into the system.

As a result of the unsatisfactory program behaviour, it was only rarely used in the scheduling process — most of the scheduling task was consequently done with pencil and paper. Regarding this situation as being even more unsatisfactory, the plant manager agreed with a proposal to solve the scheduling problem using a knowledge-based system. Main goals of this system were defined as follows .

- The system should follow the same line of reasoning the plant manager uses when generating a schedule.
- Each individual scheduling decision should be as transparent as possible to the user.
- The system should only make suggestions, "manual" changes to the proposed schedule should always be possible, the system should take care of incorporating these changes in an "optimal" way in the schedule.
- The response time should be as short as possible, a maximum of one hour was agreed upon.

In what follows, the basic idea, the implementation and first experiences with the system are described.

II. THE PROBLEM ENVIRONMENT

The fibre plant regarded here is an example of a single-step production plant : raw material enters one end of the production line, on the other end the finished product (an intermediate for textile products) leaves. The plant itself consists of several (5 to 40) such production lines with more or less different characteristics.

Planning and scheduling in this plant is completely based on customer orders : there is no need to regard storage of products nor are there problems of lot size determination. The scheduling period varies from three weeks to two months, typically there are from 300 to 1000 orders to schedule within the period. The spectrum of possible products ranges over about 12000 different combinations of product parameters, where the most prominent parameters are :

- The basic type,
- Outer (and possibly inner) diameter,
- Colour,
- Bobbin type and size.

Due to different characteristics, the set of possible production lines for furnishing an individual product is limited. A typical product may be produced on 4 to 12 distinct production lines. The exact time needed for the production of an individual product also varies with the production line used, typical variation is a factor of two.

When production proceeds from one product to another the production line has to be reconfigured to some extent depending on the parameters of the predecessor and the successor product. A main goal (from the plant managers point of view) is to keep the overall reconfiguration effort as small as possible. The orders due date is a secondary criterion from the point of view of the plant manager — as long as all orders are scheduled within the scheduling period, due dates may be neglected.

Apart from the main scheduling-goal, several other restrictions have to be satisfied by a schedule, for example :

- Regular maintenance intervals have to be scheduled for each production line.
- Randomly, production lines are unavailable for productive allocation (R & D-allocation, failures ...).
- Differently coloured versions of certain fibre types must not be produced during the same time.

III. THE IMPLEMENTATION

Since the system under construction was to follow a knowledge-based approach, we did not intend to propose an algorithmic solution, but instead decided to model the plant manager's approach to scheduling. After an initial knowledge engineering phase the following simple scheduling strategy turned out to be in use within the plant :

- Step I . Determine the possible production lines for each of the orders.
- Step II . Starting with the initial (given) allocation of production lines, look for the "most promising" successor order regarding all of the lines.
- Step III : If there are more than one equally "good" orders left from Step II, choose one of them based on secondary restrictions (due date, customer, ...).
- Step IV . Allocate the order from Step III and go back to Step II
- Step V : After all orders have been scheduled (or the capacity limit is reached), review and eventually manipulate the schedule with regard to "special cases".

Although the procedure sketched above is simple and disputable with regard to optimality, it was a procedure accepted by the plant manager and it helped to produce "good" schedules. Nevertheless — as it was done with paper and pencil — it was a time-consuming process, so it was done no more than once a month and took a couple of days of the plant managers time. Since the orders scheduled this way usually are subject to changes (in amount, type, colour ...) literally up to the last minute, the "good" schedule was changed on the fly and tended to get "worse" with every such change made.

The system constructed in response to the plant managers request in principle follows the procedure outlined above. As can easily be deduced from the above description, Steps II, III and IV of the procedure can be modeled by using a non-backtracking, forward-chaining rule-based paradigm. Step I (at least in the simplistic form it is stated) can be implemented by a database request, the central point of interest was Step V : If many and frequent changes were really made to a proposed schedule, Step V would better be viewed as being interleaved with Steps II, III and IV — resulting in a backtracking or re-scheduling strategy instead of early commitment through forward application. As our observations resulted in figures of about 5 to 10 changes per schedule, we decided to use the forward-chaining paradigm. This decision not only led to a very efficient implementation of the scheduling process — explanation of scheduling decisions has been much easier to incorporate than it would have been using a backtracking version.

The schedule is presented to the user on a graphical workstation in form of a Gantt-chart — which is accessible throughout the whole scheduling process. Whenever an order is scheduled, it immediately becomes visible on the screen and while the process is continuing in the background the user has the ability to select every individual allocation (using a pointing device) and may request an explanation for the allocation or a detailed set of informations on the selected order.

Once the proposed schedule is complete, the user may change each individual allocation in one of the following ways :

- Request, that this order should be allocated to a certain production line,
- Request, that this order is to be the direct successor of another allocated order,
- Request, that a non-scheduled order is to be the direct successor of an allocated order,
- Cancel an allocation.

Instead of directly obeying the user's request, the system gathers all of those requested changes and eventually regards them as new scheduling restrictions as soon as the user decides to start scheduling anew. This feature helps the plant manager to avoid schedules becoming worse, each time he has to do some changes. Since the whole schedule is re-done, a "good" result will again be achieved.

The most crucial decisions to be made during the scheduling process concentrate on the order selection in Steps II and III. In order to give the user a maximum of flexibility, he is allowed to edit the criteria to be used in those steps by himself. The system to this end keeps a list of "user visible rules", that allow the plant manager to define the "most promising" successor in terms of parameters of any predecessor - successor - pair. This list of "rules" is subdivided into priorized packages, the packages themselves are sorted by user-defined priorities — so the plant manager has complete control over the successor-selection. Within this rule-set the selection criteria of Step III are also incorporated, this allows the user to state the relative importance of due dates and certain other restrictions.

The system is today running on a VAX-Station using VMS, the graphical interface uses VWS (VAX Window System). The total work sums up to about 3 manyears, where about 60 to 70 percent of the effort were spent on integration aspects — the system is in close connection with existing databases and standard software components. Another 10 percent were spent for the user interface. The "AI"-part of the system has been implemented using OPS-5, the more traditional parts were done in FORTRAN, today the system consists of about 60.000 lines of code. A typical schedule proposal is generated in about 6 to 10 minutes — the time needed to complete all necessary changes is less than one hour including several recomputations of the schedule.

The system is used productively today in one plant, the rate of "manual" changes made to the proposed schedule is constantly decreasing. Two years ago we started with 20 percent accepted proposals, a few weeks ago we arrived at a ratio of 70 to 30, a realistic goal lies at about 80 to 20.

IV. CONCLUSIONS

As of today, we believe, that production planning and scheduling are typical candidates for the use of AI-techniques. In our opinion, there are several reasons, why algorithmic solutions to such problems seem unrealistic :

- Conflicting goals and an enormous amount of parameters are to be take into account.
- Goals as well as parameters often are not easily converted into numeric descriptions.
- Mathematical solutions are hard to explain.
- "On the fly" changes to schedules are hardly manageable in algorithmic attempts.
- If the environment changes, it is often hard to adopt an algorithm.

This does not mean that we blindly believe AI-techniques to be the "one and only" solution to scheduling problems, as we had to learn, a moderate mixture of AI and standard approaches seems to be a feasible way. Only if we manage to integrate small AI-parts into bigger existing organizational, hardware, and software-environments will problems successfully be solved; a purely AI-based solution would have been as much of a failure as a purely algorithmic one.

We are today continuing are work in scheduling in several directions :

- We take part in the EUREKA-project PROTOS II, where research in the direction of global, distributed and local production planning and scheduling using AI-technology is the main goal.
- We are constructing the equivalent of a "scheduling-shell" abstracting from the fibre-experiences.
- A second experiment — now regarding a more complex multiple step production — is in the queue.

Knowledge Based Scheduling in PROTOS

Jürgen Sauer
Universität Oldenburg, FB Informatik
Postfach 2503
D-2900 Oldenburg

Abstract

Within the EUREKA-project PROTOS (Prolog Tools for Building Expert Systems) [Appelrath 87] a prototype of a knowlegde based production planning system is proposed consisting of an user-interface enabling the user to plan "by hand", a heuristic planning system providing facilities to create and maintain schedules using heuristic knowledge, and a factual knowledge base of the planning area.
This paper describes a heuristic planning algorithm implemented in Prolog which aims in creating a "good" schedule of production using the experience of an expert planner. The algorithm tries to model the planning behaviour of the expert planner in using a specific planning strategy (how to create a plan step by step) and some of the heuristics applicable in guiding the search for a good solution.

1. The Scheduling Task

The task of the planning personnel is to create and maintain a schedule of production using different kinds of input information, e.g. a master schedule with given orders, information about products and resources and several conditions and constraints as well as goals to be satisfied [Fox 87, Sauer 90]. The desired solution is a plan that tries to satisfy all of the given goals and conditions. In most of the cases this will not be possible, so the planners task is to find a solution which fits best. For this he uses his knowledge and experience in guiding the search for a "good" solution.
Creating an optimal production schedule is a np-hard task. Scheduling approaches from Operations Research trying to find the optimal solution have some weak points which limit their applicability in real planning environments:

- due to the combinatory complexity of the problem area they usually work on a reduced problem model which does not fit to the real planning situation
- the algorithms are complex and the solutions and the way they are produced are difficult to understand
- most of them lack flexibility in the case of changing situations or parameters.

Intention of the knowledge based approach is to use the knowledge and experience of a production planner within a heuristic planning algorithm to build a production schedule. Several heuristic scheduling approaches have been proposed, e.g. [Fox 87, Keng 88, Liu 88], most of them try to model a specific planning behaviour using certain objectives only implicit. The first PROTOS-approach is similar to that as the heuristics/ strategies of the expert planner shall be used to create an algorithm which simulates his planning behaviour.
The planner uses an order based planning strategy where he never has to look for the whole problem area. He selects an order to be planned next and then he tries to plan the whole order at once. Orders that have been planned before usually will not be changed and so they act as new constraints for the orders to be planned afterwards.
The planning heuristic can be described schematically by the following procedure:

a master schedule with n orders is given
execute the following steps, until all orders are scheduled.

- select an order to be scheduled
- select an interval for the execution of this order
- select a production variant for the manufacturing of the product
- execute the following steps, until all steps of the variant are scheduled
  - select a step to be scheduled
  - select a possible apparatus for manufacturing the step concerned
  - if there is a conflict, try to resolve it.

For every selection the planner has to apply his knowledge, e.g. selecting 'critical' products before others (see below).
The main conflict the planner has to resolve in creating the schedule is that of overlappings (an appropriate apparatus is already occupied). One of the possible strategies to solve this conflict is (and this strategy is also used in the algorithm): first look for an alternative apparatus, if there is no one available, look for an alternative variant, if it is not possible to use the alternative variant within the given interval, look for another interval, this means shift the interval until it is possible to plan the order.

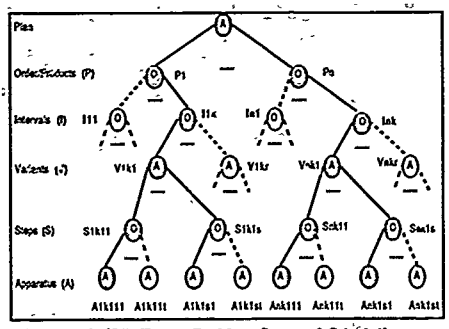


Fig. 1: AND/OR Tree - Problem Space of Scheduling

The problem area of this planning methodology can be visualized by an AND/OR-tree. The nodes of the tree contain a statement about the contribution of the children nodes to the solution and, additionally, a value that can be of different domains at each level.

A solution for the AND/OR tree is a sub-tree with:

- the root belongs to the solution
- if an AND-node belongs to the solution then all its children-nodes too
- if an OR-node belongs to the solution then exactly one of its children nodes too

Finding a solution for the scheduling problem is equivalent to finding a solution for the AND/OR tree where all the conditions are regarded
In figure 1 a solution is indicated by the solid lines.

2. The Scheduling Approach

The so called Basic-Algorithm is part of the PROTOS-planning approach which consists of

- a user-interface [Michaux 90], allowing planning "by hand" and used as man-machine-interface
- a heuristic planning system (consisting of the Basic-Algorithm and a Control-Algorithm)
- a factual knowledge base used by both of them.

Main intention of the Basic-Algorithm is the creation of a schedule of the production using the expert knowledge of the production planner. Other ones are that this creation shall be efficient and shall use the basic problem solving strategy of Prolog.
The Basic-Algorithm tries to model the planning behaviour of the planner, but as it is impossible to acquire all the knowledge and 'feelings' of the planner, only a part of his knowledge and some rules from literature are used in the planning process.

Result of the Basic-Algorithm is a schedule regarding the objectives.

- all orders (of the master schedule) are planned

- due dates should be met
- already planned orders which are e.g. planned "by hand" are regarded as hard constraints
- the production requirements (e.g. relative start dates, durations of steps, predefined apparatuses) have to be met
- no overlappings exist
- if possible the 'stem'-apparatus and -variants are used ('stem' means those variants and apparatuses which preferably should be used)
- the whole production interval for one step is occupied at once.

Meeting the due dates shall be achieved by planning 'critical' products first. A 'critical' product here means a product critical in time. A value describing the criticality is determined for every product to be planned: a product is critical if it has
- less alternative variants than others
  this is evaluated by dividing the number of alternative apparatuses by the number of steps
- to be finished before others (earliest due date)
- a smaller slack-interval than others
  where slack-interval means the distance between the earliest possible finishing date (the given start date plus the duration of the production) and the given due date
- a higher user priority than others
  this is a user given priority to prefer products.

Additionally user given weights for the priorities are used to provide the possibility to prefer one of the rules. In the current version higher weights for rules one and three are used.

The scheme of the Basic-Algorithm (fig. 2) looks much like the planning methodology presented above. Every time a choice is possible, heuristics are used.

The results of the Basic-Algorithm are presented by the user-interface or as listings.

1. check preconditions
   look for orders to be merged or splitted

2. create a list of orders sorted by 'criticality'
   heuristic: weighted combination of the four rules listed above

3. choose order
   heuristic: use the most critical product first (first of the list)
   if all orders planned -> FINISH

4. choose production interval
   heuristic: try the given start-date first, if this leads to no solution then shift adding the minimal necessary delay (this rule makes sure that a solution will be reached)
   if order planned -> 3

5. choose variant
   heuristic: try variant 0 first, then variants with increasing number (variant 0 is the 'stem' variant)
   if order planned or not plannable -> 4
   (not plannable: all variants failed for the actual interval)

6. choose step
   heuristic: try steps in reverse order (the last step first)
   if all steps planned or one step not plannable -> 5
   (not plannable: all possible apparatuses failed for the actual interval)

7. choose apparatus
   heuristic: try 'stem'-apparatus first, if not possible then try alternative apparatuses
   if apparatus free or no alternative left -> 6

Fig. 2: Scheme of Basic-Algorithm

The Basic-Algorithm is implemented in BIM-Prolog and Quintus Prolog on sun-workstations and is being tested in several selected plants. First results showing quantities of test data and runtime performance of the Basic-Algorithm are summarized in the table of figure 3.

Parameter / Plant	Plant A	Plant B
Products	33	54
Steps per product	10	3
Altern. apparatuses per step	4	3
Variants per product	2	1
Apparatuses	54	32
Orders	44	50
Planning-period (month)	6	6
Runtime (BIM-Prolog) (sec)	165	23

Fig. 3 First Results

3. Further Research

Some weak points of the prototype version of the Basic-Algorithm are:
- Not all of the necessary conditions are checked. Among them are the availability checks of personnel and raw material.
- A main point which limits the usability of the Basic-Algorithm is that it is designed for the special problem case of creating a schedule regarding the objectives and assumptions listed above. The planning heuristic (strategy) used is implemented fix. The algorithm is therefore useful only in similar problem areas.
- Another crucial point is the combination of several heurstic rules, e.g.when selecting orders. One can think of situations where the default-weights will not lead to an acceptable solution, so the weights have to be changed or in some cases new rules (e.g. regarding costs) should be added to take into account all conditions.

These topics lead to demands for a 'Control'-Algorithm which is now under development. Some of the characteristics of such a 'Control'-Algorithm are:
- Selection of heuristics for planning
  Several heuristics (strategies) for planning should be integrated in the system. The planner may choose appropriate ones or the system should propose appropriate ones.
- Using planning skeletons, adapting strategies
  Planning heuristics may be represented as skeletons using abstract operators for rules and parameters e.g. for the selection of orders. With these skeletons and appropriate rules or parameters the planner is able to build an "own" planning algorithm for a given situation.
- Integration of the Basic-Algorithm
  The Basic-Algorithm is used as one of the planning strategies.
- Selection of heuristics for replanning in specific situations
  Not only planning heurstics but also heurstics for replanning should be integrated. This leads to the possibility of a flexible and efficient reaction to unforeseen events. The selection may be done interactively or after the system proposed appropriate heuristics.
- Integration of "new" heuristics by the user
  The user may change existing or define not yet integrated heuristics using a provided definition language.

References

[Appelrath 87] Appelrath, H.-J.. "Das EUREKA-Projekt PROTOS", in. Brauer, W., Wahlster,W.: "Wissensbasierte Systeme", Springer, 1987.

[Fox 87] Fox, M. "Constraint Directed Search. A Case Study of Job-Shop Scheduling", Pitman Publishers, London, 1987.

[Keng 88] Keng, N P., Yun, D.Y., Rossi, M.: "Interaction sensitive planning system for job-shop-scheduling", in: Expert Systems and Intelligent Manufacturing, 1988.

[Liu 88] Liu, B.. "A Reinforcement Approach to Scheduling", Proc. 8th ECAI, München, 1988.

[Mchaux 90] Mchaux, G "Development of Interactive Prototypes for Production Planning & Scheduling based on Prolog", in: Appelrath/ Cremers/ Herzog:"The EUREKA-Project PROTOS", Zürich, April 1990.

[Sauer 90] Sauer, J., Appelrath, H.-J.. "Knowledge-Based Production Planning and Scheduling", in. Carnevale, M. et.al.. "Modelling the Innovation", IFIP TC7 Conference, North Holland, 1990.

KNOWLEDGE BASED PLANNING OF UNDERGROUND LIGHTING IN HARDCOAL MINING[1]

WOLFRAM BURGARD
ARMIN B. CREMERS
STEFAN LÜTTRINGHAUS
LUTZ PLÜMER
Institut für Informatik III
Universität Bonn
Römerstr. 164
D-5300 Bonn 1

AND

JUDITH GREBE
Lehrstuhl Informatik VI
Universität Dortmund
August-Schmidt-Str. 12
D-4600 Dortmund 50

AND

RAINALD GREVÉ
FRANK MÜCHER
Ruhrkohle AG
Abteilung Ergonomie
Rüttenscheider Str. 1
D-4300 Essen 1

Abstract - In this paper we present BUT, a knowledge based system for the planning of underground illumination in coal mines. The task is to find lighting configurations which are optimal from the ergonomic and economic point of view. The problem is that there is a huge number of possible configurations for lighting satisfying the ergonomic guidelines on underground illumination of coal mines developed by the European Coal and Steel Community. We present the architecture and main features of BUT which is designed to restrict the set of all possible solutions to a manageable set of ergonomically and economically acceptable or even optimal ones.

I. INTRODUCTION

Appropriate lighting minimizes the risk of accidents in hardcoal mining. In order to identify objects sufficient light intensities are required, the value of which depends on the special task to be performed. At the same time gradual changes in light intensity are desired in order to ease the accommodation of the eyes. Recently the European Coal and Steel Community has adopted guidelines which are an abstract of different research projects and define the requirements of underground illumination in coal mines [1]. For several reasons, however, it is not a trivial task to fulfill these requirements, and it is even more difficult at minimal or at least reasonable costs.

Until now, there has been no computer support for this kind of planning. The difficulties to be solved are manifold. Different tasks have different requirements on light intensity. The patterns defining these tasks cannot be characterized by simple schemes. With regard to the mine layout a lot of different scenarios have to be taken into account, which again cannot be characterized by simple schemes. Each scenario depends on the timbering, the length of the mine layout, the installed objects and equipment including their dimensions, to mention only a few parameters. For example, a stone drift may contain explosion barriers, a monorail, rails, a manway, ramps, conduits and a band conveyor as equipment. Finally, a large variety of types of explosion-proof lights is available which can be equipped with different lamps and installed in many different configurations (one or two lines etc.).

II. PLANNING UNDERGROUND ILLUMINATION

The problem to be solved can be described as follows. Given a particular mine layout including its dimensions, installed objects and equipment, identify the tasks to be performed, determine which minimal light intensities are required in different areas of that mine layout and find out which lights with which lamps can be used and where, how and in which distances they have to be fixed. An acceptable solution is one which satisfies the ergonomic requirements in each area, and an optimal solution does this for minimal costs

This is a typical planning problem. A subproblem is the calculation of the light intensity in the mine layout, given the specification of the lights, the mine layout and its equipment. This subproblem can at least be efficiently approximated with algorithmic approaches. For the planning problem, no algorithmic solution is at hand. The special characteristics of this problem suggest a knowledge based approach, since database retrieval (which lighting fulfills given specifications) combined with numerical computations (calculation of light intensity caused by a given configuration) would be too inefficient. A certain amount of reasoning is necessary to identify the areas in which tasks requiring a sufficient amount of illumination are performed and to decide where the lights should be fixed. Whereas the knowledge to identify the minimum requirements can be elicited out of the guidelines on underground illumination, the knowledge suggesting appropriate lights and mounting places has to be acquired from experts planning the lighting.

III. THE PLANNING SYSTEM

The requirements to a planning system for underground illumination are the ability to represent all necessary informations about mine lay out and to implement the guidelines and the heuristics to determine optimal mounting places Such a system also serves as a tool for the acquisition of the heuristics used by the planning experts to find ergonomically and economically almost optimal solutions, it must provide a great degree of flexibility and user-friendliness, which is a basic precondition for the tool to be accepted by the planning people Furthermore, it must be able to represent complex objects, to perform inferences with the rules stored in its knowledge base, and must provide a graphics interface to simulate different lighting configurations so that the planning people directly can test their heuristics To facilitate the implementing and test phase, an explanation tool is required allowing the access to technical documents concerning the domain. Finally it must be possible to integrate this planning system into the existing environment containing personal computer, workstations and company's data bases and software. The architecture of BUT and its interfaces to internal software systems is shown in Figure 1.

To guarantee a maximum amount of flexibility we decided to implement BUT in Prolog, since logic programming provides both a clear semantics and a practical language so that clear, efficient problem solving programs can be written within logic programming [2], [3] [4]. Furthermore, every piece of knowledge is represented explicitly,

This research carried out by the Ruhrkohle AG, Essen, is supported by the European Coal and Steal Community in the context of the fifth ergonomics program.

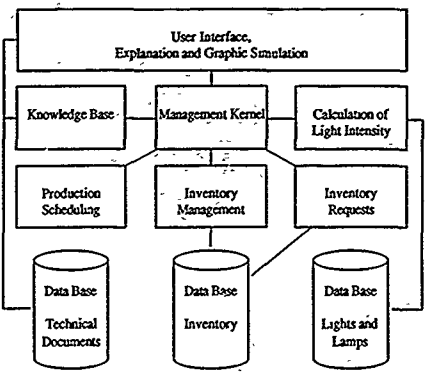


Fig. 1. The architecture of BUT

nothing remains implicit 'in fancy data or control structures'. Finally, meta-programming can be used for different extensions of logic programs providing uncertain reasoning or explanations [5].

In order to fill the knowledge base of BUT we have to represent complex underground situations and knowledge about lighting. First we have to interpret the ergonomic guidelines respecting the particular requirements of the situation. To implement them, we apply techniques described in [6]. Second we have to acquire the heuristics of the planning people. It turned out that these heuristics can easily be expressed by rules stating declarative knowledge which themselves can be translated into a form directly executable by a Prolog interpreter.

The representation of underground situations, both logically and graphically, is done in an object-oriented style, which provides abstract data types - especially for graphic objects. Following [7] the object-oriented extensions are still completely realized in standard logic programming. So we have a uniform declarative description in first-order logic yielding a neat integration of the user interface into the problem solving process.

The objects are graphically described by vector graphics. In the actual implementation BUT allows a top view including the lighting intensity and a sectional view. A special interpreter draws all objects in a particular mine layout according to the actual view. However, there is a strong relationship between the shape of an object and the installation of the lights. Some kind of spatial knowledge is used to choose optimal places where the lights have to be fixed. For example, if the lights have to be mounted to the band conveyor then its shape determines the concrete place [8].

The current prototype of BUT performs the following tasks. Following the guidelines on underground illumination in coal mines it identifies tasks to be performed and determines minimum light intensities for each mine layout. Then it tests different reasonable configurations to find out acceptable ones. An algorithmic approach is applied to optimize the distances of the lights. Subsequently all solutions are judged with respect to their costs. Figure 2 shows a screen dump displaying the result of a session in which the illumination of a stone

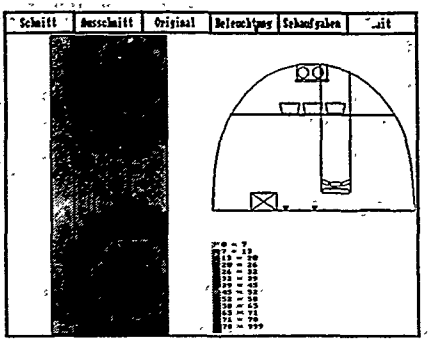


Fig. 2. Planning a stone drift with BUT

drift containing a ramp, a rail, a band conveyor, explosion barriers and conduits is planned.

IV. CONCLUSIONS

In this paper we presented a planning system called BUT for underground illumination in coal mines. To implement BUT, we integrated logic programming, object oriented programming, algorithmic approaches and graphics interfaces. BUT applies the guidelines on underground illumination in coal mines to constrain the set of possible lighting configurations. Furthermore, heuristics acquired from planning experts are used to select reasonable lights and lamps and ways to mount them in the mine layout.

V. REFERENCES

[1] Guidelines on the Ergonomics of Underground Illumination in Coal-Mines. Tech. Rept. 15, Community Ergonomics Action, European Coal and Steel Community, Luxembourg, 1990.

[2] O Keefe, R.A *The Craft of Prolog*. The MIT Press(1990)

[3] Sterling, L. Logical Levels of Problem Solving *Journal of Logic Programming 1*(1984), 151-163.

[4] Sterling, L. and Shapiro, E.Y. *The Art of Prolog: Advanced Programming Techniques*, The MIT Press(1986).

[5] Sterling, L and Beer, R D Metainterpreters for Expert System Construction *Journal of Logic Programming 6*(1989), 163-178.

[6] Arnold, G., Burgard, W., Cremers, A B., Lauer, M., and Lüttringhaus, S Ein Expertensystem zur Prüfung elektrischer Betriebsmittel für explosionsgefährdete Bereiche. In *Proceedings 2. Anwenderforum Expertensysteme*. Cremers, A.B. and Geisselhardt, W., 1988, pp. 57-64, in German.

[7] Kowalski, R. *Logic for Problem Solving*, North-Holland, Amsterdam New York Oxford(1979).

[8] Grebe, J. *Wissensakquisition für ein Expertensystem zur Planung von Beleuchtung unter Tage*, Diploma thesis, in German, University of Dortmund, 1991.

An Overview on Planning Applications in PROTOS-L

Christoph Beierle
IBM Germany, Scientific Center
Institute for Knowledge Based Systems
P.O. Box 80 08 80, D-7000 Stuttgart 80, West Germany
e-mail: BEIERLE at DSØLILOG.BITNET

Abstract: PROTOS-L is a logic programming language with types, modules, deductive database access and an object-oriented window interface. A brief overview of some planning applications developed in PROTOS-L is given.

1 Introduction

The EUREKA project PROTOS [Appelrath *et al.*, 1990] is concerned with the development of logic programming tools and their applications in building expert systems for planning applications. Current partners in PROTOS are BIM, Hoechst, IBM Germany, Sandos, and the Universities of Bonn and Oldenburg.

Within PROTOS the logic programming language PROTOS-L ([Beierle, 1989], [Böttcher, 1990]) has been developed. PROTOS-L has types, modules, deductive database access ([Meyer, 1989]) and an object-oriented interface to OSF/Motif ([Jasper, 1991], [Scheuk, 1991]). The purpose of this paper is to give a brief overview on some planning applications that have been developed in PROTOS-L.

2 Map Colouring

For illustration purposes we have implemented a version of the well known map colouring problem. Given a set of countries together with their neighbouring relations, each of these countries is painted with one of four colours (say: green, red, yellow, blue) such that no neighbour countries have the same colour. For this problem domain the facilities of PROTOS-L of constraining variables to subtypes were exploited. For instance, given a map of Europe, assume that Switzerland gets the colour red. Immediately, all neighbours of Switzerland are constrained to be not red, i.e. their colour must be in {green, yellow, blue} only. If Germany then is assigned the colour green, e.g. France's colour value must be in {yellow, blue}. This propagation of constraints significantly reduces the search space. When the constraints yield an empty set of possible colours for any country this failure is detected as early as possible, and not only when the colouring of that country is tried. To give an example, the standard backtracking solution in Prolog used 126 backtracking steps for colouring a map of Europe, whereas the PROTOS-L solution using type constraints as sketched above found the first solution without a single backtracking step. This example illustrates that a similar benefit can be gained by constraint propagation techniques in production planning (see e.g. [Dincbas et al., 1988]).

3 IC Routing

Finding a route in a graph under certain conditions is another classical planning situation. A particular instance of this problem domain is solved in the IC route planning system that has been developed in PROTOS-L.

A major part of the German railway system is the InterCity net that connects over 50 cities in Germany. In our system, the direct city-to-city IC connections are stored in a relational database, and the problem is to find a good IC connection between two cities given by the user, possibly with changing trains. Besides the two cities, a time T must be given and optimisation criteria like:

- arrival as soon as possible, depature at T or after T
- departure as late as possible, arrival at T or before T

A high level user interface has been developed on an IBM RS/6000 workstation using PROTOS L's object-oriented interface to X Windows and to OSF/Motif ([Scheuk, 1991]). The departure and arrival cities can be selected by mouse clicks, and there are menus for the different options. The system finds the *optimal* connection under the given options.

Some of the experiences we made with the IC route planning system were:

- An efficient and intelligent close connection to an external relational database is vital in such applications. A deductive database component allowing also recursive queries as in PROTOS-L is useful for easy formulations of path finding problems and can even avoid termination problems in case of cycles in the data ([Meyer, 1989]). The database access strategies have to be accomodated to the set-oriented evaluation in databases, which is opposite to the backtracking method in logic programming.
- The particular transparent database access realised in PROTOS-L made it very easy to move from a prototypical test implementation with only a few cities and some connections in main memory to the full database version. This rapid prototyping aspect of the database access has also been exploited in the PPS applications to be discussed below.
- The system is very flexible w.r.t. what is an optimal connection. For instance, instead of the fastest connection one could easily take the connection with the smallest number of changes as the optimal one.
- High-level user interfaces are not only useful for applications requiring a great amount of user interaction, because applications as this one can be used with virtually no training effort at all.

4 Multi-step production planning

The main application domain of the PROTOS project is production planning and scheduling. A PPS problem description provided by the PROTOS partner Sandos ([Sauer *et al.*, 1989]) occurs in the area of pharmaceutical production planning:

- There is a set of orders to be produced. Each order specifies a product, an amount, and a due date. Orders are also assigned priorities.
- Each product can be produced in a number of variants. Each variant consists of a number of production steps, and each production step can be carried out on a number of different machines.
- Machines have different characteristic features, not every step of a variant can be produced on every machine. But the machines are multi-purpose machines as they can do many different jobs.
- Production constraints have to be obeyed. For instance, the individual production steps of a given variant must be executed in a predefined sequence, without any time delay between them.

For this multi-step production planning system we have developed different strategies:

1. Guided by Expert: As soon as an order with highest priority can not be produced before its due date the planning process stops in order to allow the planning expert to decide what to do about this situation.
2. Maximal Planning: If higher priority orders can not be produced before their due date any more, the planning process still plans a maximal number of next level priority orders that can be produced before their due date.
3. Delay Orders: If an order with highest priority can not be produced before its due date it is nevertheless put into the plan. An information is given to the user how much such orders will be delayed after their due date.
4. Combination of (2) and (3): As many orders as possible are planned, starting with highest priority orders first and allowing exceeding of due dates where necessary.

According to experiences made by Sandos it was tried to solve this multi step PPS problem with classical OR methods, but without success. The four different planning strategies sketched above do not guarantee an optimal plan in a mathematical sense, but human planning expert can use them and experiment with them to see what effects they have for a particular set of orders. Among the experiences drawn from this application was that it was very helpful to have the system implemented in a high level logic programming language, for instance, the changing of any parameters and strategies in the planning system was much easier.

Also in this application there was a smooth transition from a small test version using only main memory to the full version with the external database relations containing the orders information, the production requirements with their variants, steps and machine alternatives, etc. The typing concept helped to discover many programming errors already in an early phase of program development which might have otherwise caused much more difficulties in locating and eliminating them. However, the first versions of this planning system were developed while the OSF/Motif interface of PROTOS L was not yet available. Thus, to work with the system without a window-based user interface is by far not as comfortable as with such an interface as it has been developed for the PPS application presented in the next section.

5 Single Step Production Planning

The following PPS application deals with fiber production as it occurs at the PROTOS partner Hoechst AG. The problem is to assign a set of orders (typically between 100 and 500 per month) to a number of production lines (say, for instance, 15 lines). Once the production of an order has been started on a production line it will also be finished on that line before anything else is produced on the same line. Thus, this problem is an instance of a single step production planning problem.
The main characteristics of the problem situation are:

- Each order specifies a certain type of fiber, the amount to be produced, a bobbin type, whether the fiber should have a profile or not, a diameter, and a due date. Furthermore, it may specify a certain colour or it may specify colourless.

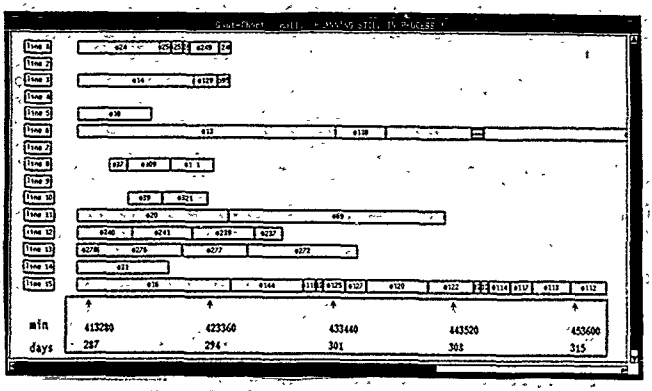


Figure 1: Gantt Chart produced by the HoPla system

- The production lines may differ in their characteristics. There are slow and fast production lines, each line can produce only a subset of all types of fibers, etc. From the production line characteristics it is possible to determine which order can be produced on which production line. This may depend on all parameters of an order like its fiber type, its bobbin type, its diameter, etc.
- Producing an order S after an order P on the same line will cause a certain amount of resetting costs. The amount depends on the line and P's and S's fiber type, colour, diameter, and bobbin type. Thus, these costs can only be determined once a production line has been selected for these orders
- Apart from the production restrictions already mentioned, there is one major restriction which is called the colour constraint. There are special fiber types which form a subset of all possible fiber types. The constraint is that no two orders for these special types can be produced at the same time on any of the production lines if the orders specify two different colours. Therefore, orders for Coloured Special fiber types (called CS orders, for short) must be handled especially careful in the planning process. For instance, a green CS-order running on line 4 from time T1 to time T2 will prohibit the production of any blue, yellow, red etc. CS-order on any other line between T1 and T2

All data mentioned so far is stored in an external database. The goal of the planning process is to find a schedule such that all orders are produced in time and all restrictions and constraints are satisfied. Since the resetting costs can be rather high (e.g. when changing from one colour to another) a important objective is to minimize the overall resetting costs of a plan.

According to the experiences of Hoechst, also this PPS could not be solved successfully with linear programming techniques. The computation time was too long (several hours), the resulting plan could not be modified (which had to be done frequently due to modified orders), and the plan often did not reflect additional constraints the human experts knew and obeyed but that were not reflected in the program (which would also require a plan modification). Therefore, a knowledge based approach was taken at Hoechst ([Jachemich, 1990]) where the planning knowledge of the experts could be expressed directly. Certain heuristics to select a "best next" order cut down the search space significantly. For instance, there are rules like

- A best next order should have the same parameters as its predecessor.
- If one has to switch colours or diameters rather take the same colour and switch the diameter.
- If one has to switch colours, it is better to switch from a lighter colour to a darker one.
- If one has to switch diameters, it is better to switch from a smaller to a greater one.

Each of these rules expresses a planner's knowledge how to keep the resetting costs low. A system based on such rules has been developed at Hoechst using OPS5, and this approach has also been the base for the PROTOS-L planning system called HoPla (Hoechst Planning System). In addition to the data listed above, the planner gives a time interval for the planning period, and a list of colour intervals. Each of these colour intervals specifies a sub-interval of the planning interval in which the system may plan CS-orders of that colour. The HoPla system generates a Gantt chart displaying the produced plan that is coloured according to the orders' colours (Fig. 1). After an order has been placed into the plan it is immediately visualized in the Gantt chart. Already during the still ongoing planning process, every order that has been planned so far can be clicked with the mouse. A new window appears and shows the details of the clicked order, the resetting costs it has caused, and the reason why it has been placed into this position in the plan. Thus, an explanation facility is available for the planner, enabling him to modify or add, if necessary, any constraints (like placing one order directly after another one) or any planning rules.

The experiences with the HoPla system were similar as with the Sandox planning system described above as far as the database access and the use of a typed logic programming language are concerned. The modification and addition of the knowledge base is quite easy. In addition, the window based user interface made it much easier to work with the system, e.g. to see what effects a change of the colour intervals or changes or additions of planning rules have.

The typical time to produce a plan with 300 orders is approximately 10 to 15 minutes in our current implementation. Optimizing the run time of the system is possible, but one has to be careful not to trade run time efficiency for flexibility while still in a prototype phase. A detailed comparison of the HopPla system with the system developed at Hoechst and implemented in OPS5 has still to be carried out. Currently, we are planning to add further functionalities to the system which are in the focus of the second phase of the PROTOS project, namely capacity information for production lines, ware houses, etc.

Acknowledgments: The development of PROTOS-L and its applications has been a major effort which would not have been possible without the work and support of many researchers, guest scientists, students, and our colleagues from our project partners, it would not be possible to mention all of them here. However, apart from those that have already been cited explicitly above I would like to thank R. Hauser, P. Sanders, and G. Urban who made major contributions to the PPS applications described in the last two sections.

References

[Appelrath *et al.*, 1990] H.-J. Appelrath, A. B. Cremers, and O. Herzog. *The Eureka Project PROTOS*. Stuttgart, 1990.

[Beierle, 1989] C. Beierle. Types, modules and databases in the logic programming language PROTOS-L. In K. H. Blässus, U. Heidtstück, and C.-R. Rollinger, editors, *Sorts and Types for Artificial Intelligence*, Springer Verlag, Berlin, Heidelberg, New York, 1989.

[Böttcher, 1990] S. Böttcher. A tool kit for knowledge based production planning systems. In *Proc. Int. Conference on Database and Expert System Applications*, Vienna, 1990.

[Dincbas *et al.*, 1988] M. Dincbas, H. Simonis, and P. Van Hentenryck. Solving a cutting stock problem in contraint logic programming. In *Proceedings Fifth International Conference on Logic Programming*, Seattle, WA, August 1988.

[Jachemich, 1990] J. Jachemich. *EXAMPL*. Working Paper, Hoechst AG, 1990.

[Jasper, 1991] H. Jasper. A logic based programming environment for interactive applications. In *Proc. Human Computer Interaction International*, Stuttgart, 1991 (to appear).

[Meyer, 1989] G. Meyer. *Rule Evaluation on Databases in the PROTOS-L System*. Diplomarbeit Nr. 630, Universität Stuttgart und IBM Deutschland GmbH, Stuttgart, December 1989. (in German)

[Sauer *et al.*, 1989] J. Sauer, G. Michaux, and L. Slahor. Wissensbasierte Feinplanung in PROTOS. In W. Brauer and C. Freska, editors, *Proceedings GI Kongreß Wissensbasierte Systeme*, Springer-Verlag, 1989.

[Schenk, 1991] M. Schenk. *Graphical User Interface for the PROTOS-L System*. Diplomarbeit Nr. 763, Universität Stuttgart und IBM Deutschland GmbH, Stuttgart, January 1991. (in German)

PPS Views: Scheduling and Replanning

G. Michaux
BIM
Kwikstraat 4 - 3078 Everberg (B)

Summary

For many years, planning and scheduling applications are the field of intensive research and development in operation research as well as in Artificial Intelligence, in particular through the EUREKA project PROTOS. We give main conclusions drawn from our experience in this field. After a brief description of the production planning and scheduling process, we give some outlines of the integrated environments we developed. In spite of their success in real world PPS applications, such tools including scheduling algorithms and graphical facilities were too restrictive. People in charge of PPS applications let appear new requirements, especially with respect to replanning and reasoning issues. The so-called descriptive view we propose to approach these requirements is presented in the last part of this paper. General conclusions deal with the usefulness of prototypes within the framework of PPS applications.

1. Introduction

Artificial Intelligence techniques currently present a growing interest in the field of production planning and scheduling (PPS). Numerous developments have been achieved in order to obtain knowledge based systems related to PPS applications [Oliff 88, AIESM 90, PROTOS 88, PROTOS 90]. These systems aim at complementing the functionalities and scope offered by existing MRPII systems. These deal with long term planning and manage information such as sales orders, work orders, bills of materials, as well as inventory status. These systems however present major restrictions: they are heavy environments and not well adapted to the daily changing conditions of the busy real world life. Furthermore, the actual capacity of the plants is not taken into account. There is definitely a gap between classical MRPII systems and the daily practice of the plan managers. This paper aims at presenting results of the experience we acquired in practical applications in the field of job shop scheduling, discrete parts manufacturing as well as of scientific experiments planning. Commercial contracts and the EUREKA project PROTOS[2] enabled us to propose methodological views on the concerned scheduling problems as well as to build executable models [Michaux 90]. Main conclusions of these developments first relate to the scheduling process, i.e. how the resources of shop floors are allocated to production steps. We then mention the enlargement of this view - such as proposed in the second phase of PROTOS - and show how reasoning and replanning issues could be integrated. General conclusions emphasize the development of Prolog based explorative prototypes as a powerful mean to work out the many facets of complex applications such as PPS ones.

2. Scheduling Issue

2.1 Problem overview

The PROTOS problem we refer to in this paper concerns the scheduling of a few dozen work orders which might be achieved according to one production routine selected from a set of routines, themselves composed of 1 to 20 steps. The application deals with a multi-purpose apparatus environment. There exists no commercial shell to manage or support the following activities which are performed in many PPS areas, even within the restricted framework of one single shop floor:

- order acceptance: all the work orders must not be necessarily planned. Because of the numerous factors conditioning the PPS activities, the plan managers have the responsibility to select which orders they want to schedule. They can also suggest modifications of these orders;
- industrial engineering: the organizational constraints according to which the resources are involved in the production recipes cannot be considered as static information. They might be regularly updated to take into account the continuously evolving techniques of production as well as exceptional opportunities or impossibilities to use particular resources. Support is needed to ensure the consistency of the knowledge base implementing the organizational constraints;
- scheduling: classical operations research techniques failed to find solutions for real world applications because the combinatorial explosion phenomena is tremendous in these applications. Order-oriented algorithms perform breadth-first search in the decisional tree corresponding to the successive selections of appropriate routes, production steps, resources and time windows [PROTOS 90] As many scheduling algorithms, they however rely on semi-empirical heuristics which refer to some local property of the PPS problem;
- decision-making: beyond the order acceptance, the plan managers may also decide to add additional constraints, whatever the reasons therefore - internal or external to their shop floor For example they may want to impose a given production route for one particular product within a given time window. The constraints as well as the goals evolve dynamically according to the current PPS context which is changing and evaluated on run time by the user.

2.2 Executable Environment

Providing the plan manager with efficient support requires to integrate these different facets into one integrated environment. Part of the PROTOS project was to build one explorative prototype reflecting this approach [Michaux 90]. This prototype was completely implemented in Prolog. It was actually in use in several dye stuff plants. It is a multi-windowing system capable of integrating some scheduling skills:

- work load and application management: the work orders can be retrieved and selected, and their properties updated after some rule checking -e.g., the duration of production has to fulfill some default rule based on the recipes but this rule can be overridden if the user confirms a different option. Besides, the order-oriented scheduling algorithms can be launched;
- shop floor model: the production routes are displayed, with highlighting features in order to draw user's attention on the selected resources or on the conflicts. Besides, these routes may be interactively created or updated at run time;
- plan control: its first purpose is to allow the user making a visual evaluation of the current plan. The allocations -for production or maintenance- can be seen through different filters, built interactively at run time. Plan control enables to simulate the planning activities of the plan managers: these can select either work orders, production steps or allocations. Scheduling actions can be applied to these selected items in order to modify the existing schedule, e.g. through the simple shift of Gantt charts or by decreasing the normal duration of production. The conflicts are notified in a message window;
- planning survey: providing the plan mangers with one additional view of the plan.

Many others facets could have been added, such as the integration of shop floor tracking or yet the generation of diagnosis reports. The existing prototype however definitely allowed the experts to simulate the activities they performed in their daily planning tasks As they felt very comfortable with this tool, this development enabled to face the well known knowledge acquisition bottleneck [Prerau 87].

3. Replanning Issue

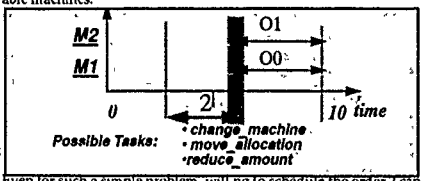
3.1 Hierarchical PPS View

Planning generally deals with the sequence of tasks leading to the achievement of orders while scheduling relates to time assignments to such tasks. Actually, the distinction between both concepts is practically rather fuzzy. Moreover, plan mangers are concerned by both re-planning and re-scheduling activities. Taking into account new orders or yet events occurring in the resources park obviously involves to re-schedule some parts of the existing plan. Adequacy of solutions brought by scheduling algorithms may be questioned here. Even with dynamical priorities, introduced to select the most "critical" scheduling operation [Keng88], these algorithms just perform a rough and simple management of scheduling actions: these always attach time frames to resources Moreover they must still be applied to a predefined set of work orders. Qualitatively, the plan managers do more: they make judgement first on the scheduling problem, and then on the appropriate actions on basis of general assessments on the current plan. Rather than focusing on the chances for particular scheduling actions to succeed (or fail), they use a mixed reasoning. At the same time, they consider the critical production steps and also estimate the opportunity to plan which orders could be scheduled partially, completely or not at all They play with more flexible planning constructions which tolerate both incompleteness or even inconsistency. The local scheduling problem is actually an open one, in continuous exchange with the logistic departments and the plan mangers who cooperate in order to define the master schedule. More than just solving scheduling problem at the level of individual plants, the planning and scheduling process continuously evolves according to an hierarchical decision schema. Such as proposed in PROTOS [PROTOS 90], "global planning" would cope with the information normally managed by the MRPII systems but integrating moreover the current -changing- capacity of the plants, while "distributed scheduling" would specify, for the different production plants, the work orders, i.e. the input to the "local scheduling" activity of these plants.

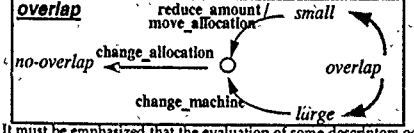
3.2 PPS Descriptive View

Mastering this local scheduling -and planning- activity asks for being able to assess a few general statements on the PPS problem. These statements should help in focusing on the targeted goal by restricting the number of possible tasks and constraints to be considered at the most detailed level. Managing such statements, without hampering the solving process at this level, is the concern of planning techniques These need descriptors to qualify the state of

the concerned application. In real world PPS applications, the working basis on which the plan managers assess their decisions does not always take into account all constraints. In building a plan, the plan managers may very well focus on the reservation of apparatus for some specific operations. On the opposite, for other operations, they might check the availability of resources only at the last minute and even allow temporary conflicts. Such a case occurred with some widely performed operations: the plan managers first assume they will always find a solution. In the practical applications we encountered, we could more generally distinguish essentially relaxable, i.e. very soft, constraints from hard constraints. Fulfilling the latter should be automatically solved by the system: the soft ones could be ignored but taken later into account in an interactive process between the user and the system. The user would be guided through dialogs in searching solutions for the soft constraints, on basis of his evaluation of the current context of the shop floor as well as suggestions made by the system. The scheduling actions which can be applied to solve a conflict depend on the characteristics of these conflicts. It is the purpose of the descriptors both to achieve such characterization and to filter the appropriate scheduling actions. Integrated into a hierarchical planning framework, such descriptors would enable to avoid dealing directly with the innumerous detailed allocations of the PPS world but well to reason on some descriptions of the schedule or of conflicts. This is the reason why this approach is called "descriptive view".
The figure shown below presents a very simplified PPS case where one order, O2, has to be scheduled on one of two possible apparatus, M1 or M2, in a context where 2 single-steps work orders, O0 and O1, are already scheduled. In order to schedule this order, one is forced to create conflicts with existing allocations for both available machines.



Even for such a simple problem, willing to schedule the order 2 can give rise to numerous solving paths. Scheduling actions could modify the existing allocations - and allow to insert the new order without conflict- but these scheduling actions may well introduce new conflicts: for example, moving the order 1 -using same machine as order 2 - could violate a precedence constraint on the orders 0 and 1. Hierarchical planning techniques have been proved to constitute the basis to guide the reasoning of planning experiments in the field of molecular genetics [Stefik 88]. The descriptive view we propose aims at organizing the knowledge according to a hierarchical representation enabling to reason on the possible scheduling tasks in order to solve well specified PPS problems. Willing to achieve one goal -here, to remove one overlapping conflict-might ask for considering a large number of allocations (PPS world). This number is however already limited through the recipes (shop floor model). Introducing the descriptive layer dramatically reduces the possible scheduling actions according to which this goal could be achieved, such as illustrated on the next figure.



It must be emphasized that the evaluation of some descriptors occurring in a solving process, for example the importance of an overlapping conflict - small vs. large- might be prompted to the user. A lot of detailed features indeed determine the functioning of the physical resources, and sometimes may dramatically influence the decisions taken by the plan managers. One should consider that all production requirements might not necessarily be known by the system for example, intermediate products might be transferred into another production center where other resources are available- but such a transfer depends on several specific circumstances quantity, nature of the product. The goal is to support dialogs between the user and the system and to integrate into the solving process some judgements the user might declare at run time.

4. Conclusion

Since the production planning and scheduling process is a complex and ill-structured domain, many conclusions could be drawn from the applications we encountered. We want here to point out two major ones. First, the usefulness of explorative prototyping has been proved in the field of Artificial Intelligence. In dynamical contexts such as encountered in large business companies, we again confirmed that explorative and quick prototyping are required in order to capture, and even to create, the domain knowledge. It must be stressed that Prolog, a logic programming language including an inference engine and defining a rule syntax, can play a major role in such contexts. It indeed provides, in one unified corpus, a high level, conceptually sound, readable and executable specification of the problem. Finally, for many years, scheduling has received a lot of attention from operations research. However, because of the NP-hard nature of the real world problems, and though achieving performant algorithms still remain an intensive research field, the major concern of plan managers appears more and more as being able to reason on their daily scheduling practice The experts now expect tools capable of guiding them in the way to solve specific PPS problems through interactive dialogs with the system.

References

[AIESM 89] Proc. of the 1st International Conference on "Artificial Intelligence and Expert Systems in Manufacturing", 20-21 March 1990, London, UK

[Keng88] "Interaction-Sensitive Planning System for Job-Shop Scheduling", N.P. Keng, D.Y.Y. Yun and M. Rossi, in [Oliff 88], pp. 57-69

[Michaux 90a] Michaux, G.: "Développement d' environnements intégrésd'aide à la planification basé sur Prolog", in "Les utilisations industrielles du langage Prolog", pp.49 - 59, AFCET, Paris, Avril 90.

[Michaux 90b] Michaux G, de Zegher I, Slahor L., "Building a scheduling system with Prolog based tools", Les Systèmes Experts et leurs Applications - Conference Sectorielle "IA,Industries Agro-Alimentaires, Biologie, Chimie et Pharmacie", Avignon - 31 mai 1990

[Oliff 88] "Expert Systems and Intelligent Manufacturing", Proc. of the 2d Int. Conf on ES and the Leading Edge in Production Planning and control-M.D. Oliff Ed.- North Holland, 1988

[Prerau 87] Prerau, D.S., Knowledge Acquisition in the Development of a Large Expert System, AI Magazine, 43-52, Summer 1987.

[PROTOS 88] "PROTOS -First PROTOS " H.-J. Appelrath - A.B Cremers - O Herzog Ed., Sandoz AG, Abt. Methoden & Modelle, Postfach, CH-4002 Basel - December 1988

[PROTOS 90] "The EUREKA Project PROTOS", H.-J. Appelrath - A.B. Cremers - O Herzog Ed., IBM Deutschland GmbH - Wissenschaftliches Zentrum - Institut fur Wissenbasierte Systeme - Postfach 800880 - D 7000 Stuttgart 80 - Apnl 9 1990

[Stefik 88a] Stefik M. "Planning with Constraints (MOLGEN)" Part 1, in *Artifical Intelligence* 16,111-140,1981

[Stefik 88b] ibid, Part 2 pp. 140-169,1981

1 MRPII = Manufacturing Resources Planning

2. Partners of the EUREKA PROJECT EU 56 PROTOS - Prolog Tools for Building Expert Systems were BIM (Belgium), ETH Zürich (Switzerland), IBM Deutschland GmbH (Germany), Sandoz AG (Switzerland), Union Bank of Switzerland (Switzerland), Universität Dortmund (Germany), Universität Oldenburg (Germany). Since 1st Apnl 1991, PROTOS entered into its second phase, more oriented to PPS applications and with BIM, IBM Deutschland, Sandoz, Universität Dortmund, Universität Oldenburg and HOECHTS as partners.

A RE-EXAMINATION OF THE FISHER HYPOTHESIS

Unit Erol
Department of Economics
Bilkent University
Ankara, Turkey

Thomas R. Colledge, Jr.
Decision Sciences Department
George Mason University
Fairfax, VA 22030 USA

James A. Richardson
Department of Economics
Louisiana State University
Baton Rouge, LA 70803 USA

The Fisher Hypothesis and Uncertainty

The Fisher Hypothesis has been the subject of extensive empirical and theoretical analyses since it was proposed by Fisher.[2,3] The Fisher hypothesis assumes that nominal interest rates are related to anticipated price changes, leaving the real interest rate constant. In other words, the nominal interest rate can be expressed as the sum of two components:

$$R = r + a \tag{1}$$

where R is the nominal interest rate, r is the real interest rate, and a is anticipated inflation. Classical tests of the hypothesis replace unobservable anticipated inflation in equation (1) with a weighted lag distribution of past price increases, ($\sum v_i \, P_t / P_{t-i}$), where the v_i are the distributed lag weights. The Fisher hypothesis, as a statistically testable proposition, takes the following form:

$$R = \alpha + \beta \sum v_i \frac{P_t}{P_{t-i}} + \varepsilon_i \tag{2}$$

where ε_i is a random error term.

Equation (2) is a dynamic impulse response equation. For parameter estimation, the lag structure is truncated, and the following equation is considered:

$$R_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_m X_{t-m} + \varepsilon_t \tag{3}$$

with $\sum_{i=0}^{m} \beta_i < \infty$ and X_t is the rate of change in the price level. One important concern about equation (3) is the fact that this form precludes the existence of feedback between interest rates and prices.

A second aspect of the hypothesis that deserves further attention is theoretical in nature. The theoretical form of the Fisher hypothesis can be derived from a simple choice-theoretic model. Market participants must select between present and future consumption claims in an intertemporal model. The details of this approach are presented by Hirshleifer.[4] This model is not discussed in this short version of the paper, but a longer version is available from the authors by request. We proceed in this paper to a direct extension of the framework.

If we reconsider the subject within the theoretical framework of Debreu[1] and Hirshleifer[4], market agents must consider the possible states of the world upon which the receipt of future consumption claims depend. Let $c_A^1, c_B^1, \ldots, c_N^1$ be a set of alternative consumption claims for future time period 1 and $\pi_A^1, \pi_B^1, \ldots, \pi_N^1$ the respective subjective probabilities assigned to the alternative claims at time 1. This may easily be generalized to N periods with the same set of respective probabilities, assuming the independence of the assigned probabilities with respect to time. The behavioral rule that governs the decisions of market agents in this context is the maximation of expected utility

$$\pi_A^A v(c_0^A + \cdots + c_1^A) + \cdots + \pi_N^N v(c_0^N, \ldots, c_1^N)$$

where v is the utility function of a representative individual. Assuming only one future time period (1 versus present), and two alternative states A and B; the maximization of expected utility under a given budget constraint yields the following equilibrium conditions:

$$-\frac{\theta_{1A}}{\theta_0} = -\frac{1}{1 + r_1^A} =$$

$$\frac{\pi_A \dfrac{\partial v(c_0, c_{1A})}{\partial c_0} + \pi_B \dfrac{\partial v(c_0, c_{1B})}{\partial c_0}}{\pi_B \dfrac{\partial v(c_0, c_{1B})}{\partial c_{1A}}} \tag{4}$$

$$-\frac{\theta_{1B}}{\theta_0} = -\frac{1}{1 + r_1^B} =$$

$$\frac{\pi_A \dfrac{\partial v(c_0, c_{1A})}{\partial c_0} + \pi_B \dfrac{\partial v(c_0, c_{1B})}{\partial c_0}}{\pi_B \dfrac{\partial v(c_0, c_{1B})}{\partial c_{1B}}} \tag{5}$$

where θ_{1A} and θ_{1B} are real future prices associated with alternatives A and B respectively, while r_1^A and r_1^B are risky real rates for alternatives A and B. Equations (4) and (5) imply that real interest rates are, in a world of uncertainty, also functions of the probabilities assigned to future events and the market participants' utility functions. The importance of this result, with respect to the Fisher hypothesis, is the following.

We can, more realistically, write $\pi_A(A)$, $\pi_B(A)$, $\ldots$, $\pi_N(A)$ where A is a general set of information that is utilized by a representative market agent in the current period to assess the possible states of the world. This form implies that the probabilities assigned to future events are based on all currently available information. The incorporation of a new signal can affect the subjective probabilities of a representative agent and hence, through equations (4) and (5), the real interest rate. It is clear that the real interest rate may be subject to change, even over a short time period, given this context and a modern

economy with well-developed mass media and relatively efficient markets. It is evident that the Fisher hypothesis is implicitly restricted to the utilization of information concerning only the future depreciation of the monetary unit in a money economy.

Thus the information set implied by equation (3) is $A_F = (\bar{X})$ where $(\bar{X})$ is the set of present and past inflation rates. The information set suggested by equations (4) and (5) is $A_C = (\bar{X}^{-1}\lambda)$ where λ is the information subset that includes other types of information (excluding inflation) in the assessment of future states of the world.

Furthermore, these two approaches yield two different statistical hypotheses that can be tested. The Fisher hypothesis implies equation (3) as a statistically testable vector-autoregressive or vector-ARIMA form. The disturbance term is an independent identically distributed white noise process which is assumed to be uncorrelated with both the explanatory and dependent variables.

The statistical form implied by equations (4) and (5) is more subtle and requires additional explanation. Any new information λ_i, where λ_i is an element of set λ, alters the overall information structure, A. Consequently the probabilities π_A, π_B, . . ., π_N are altered since they are functions of A. A change in the probabilities changes the real price of future consumption claims in terms of present consumption claims, and therefore the real interest rate changes through equations (4) and (5). Real bond prices (inversely related to interest rates) will tend to increase since bonds are assets that give claim on future consumption. If the situation is characterized by normal capacity utilization, efficient processing of information, continuous past increases in the price level, and a two-asset (bond-commodity) market; any new information, λ_i, that increases the premium on present consumption versus the future consumption would theoretically have a positive effect on the bond supply and a negative effect on the real demand for bonds. This would exert downward pressure on the real interest rate. Conversely, the increase in the premium on present consumption claims will increase current (real) aggregate demand, thus putting upward pressure on the price level and "fueling-up" the inflationary process if the increase in real output cannot totally absorb the increase in current aggregate demand. Accordingly, new information λ_i incorporated in the set A may simultaneously lead to an increase in both the inflation and real interest rates. The observable nominal interest rate will then increase due to the increase in real rate.

The real causal factor in this scenario is the incorporation of new information in a representative individual's information set, A. The λ_i are usually incorporated in the disturbance term since these factors are usually nonquantifiable. The implication is clear. Any regression of nominal interest rates on the inflation rate will be affected by the correlation of both variables with the disturbance term. This renders efficient estimation impossible. If the adjustment of both the commodity and bond markets to new information is rapid, under the assumption of efficient markets; inflation and the nominal interest rate will be observed contemporaneously (or nearly contemporaneously if there is a slight adjustment lag) correlated with the disturbance term.

The remaining issue is the following. Can this theoretical proposition be statistically verified? The conventional econometric approach, using a reduced form or a simultaneous equation model, can be ineffective given the problems of correctly modeling the information set A and specifying the correct a priori restrictions. There are, however, techniques that are designed for estimating parameters in multiple equations given time series data.

Multiple Time Series Models

Multivariate models are used to represent two or more multivariate stochastic processes. The general form of the model can be written as follows:[5]

$$\phi(B)\bar{Z}_t = \theta(B)\bar{a}_t \tag{6}$$

where

$$\phi(B) = I - \phi_1 B - \ldots - \phi_p B^p, \tag{7}$$

$$\theta(B) = I - \theta_1 B - \ldots - \theta_Q B^Q. \tag{8}$$

In the above formulation $\phi(B)$ and $\theta(B)$ are kxk matrices of autoregressive and moving average operators respectively. Both $\phi(B)$ and $\theta(B)$ are matrix polynomials in the backward shift operator (B) of degree P and Q respectively. The vector $\bar{Z}_t$ contains the variables to be included in the multivariate stochastic analysis.[6] Finally, $\bar{a}_t$ is a vector of random, independent, identically normally distributed shocks with mean zero and covariance matrix Σ.

Parameter Estimation in the Multivariate Model

We initially estimated a multivariate model with the nominal interest rate and inflation rate as variables for the 1959-1983 period. A total of 292 monthly observations were used. The data for interest rates are 90-day monthly treasury-bill yields from the Federal Reserve Bulletin. The inflation rate series is constructed by taking the first-difference of the natural logarithm of the consumer price index (CPI).

The identification and estimation stages suggested the following model

$$(I-\phi_6 B^6)\bar{Z}_{1t} = (I - \theta_1 B^1 - \theta_2 B^2)\bar{a}_t \tag{9}$$

as the most adequate and parsimonious model on the basis of diagnostic checking. The 2x1 vector, $\bar{Z}$, contains the transformed inflation and nominal interest rate sereis, and $\bar{a}$ is a 2 x 1 vector of independently distributed random shocks. The parameters are contained in the 2x2 matrices ϕ_6, θ_1, and θ_2. The residual diagnostic checking confirms that the model is adequate; that is, the model transforms all the sample cross-correlation matrices into realizations of white noise processes.

The generating mechanisms for the data are more easily seen using the following explicit form:

$$(1+.216B^6)X_t = (1-.763B)a_{1t} + .004Ba_{2t} \tag{10}$$

$$(1+.170B^6)R_t = \tag{11}$$

$$(5.174B - 3.53B^2)a_{1t} + .(1+.412B)\,a_{2t}$$

In equations (10) and (11) R_t is the first difference of the logarithm of nominal interest rates, X_t is the first difference of the logarithm of the inflation rate, a_{1t} is the innovation driving the inflation rate, and a_{2t} is the innovation driving nominal interest rates.

The model is dominated by the moving average terms. The inflation rate series is driven by the innovations of the interest rate with a short lag, and the interest rate is driven by lagged innovations of the inflation rate. The series show strong seasonal autocorrelation. The distributed lag of the actual inflation rate does not have a significant effect on the current nominal interest rate, suggesting that the information relating to past inflation rate changes may not have a very signficant role in the overall information set A. The overall results are strongly suggestive of the alternative hypothesis put forward in the previous Section since it is the inflation rate innovations that explain nominal interest rate rather than the inflation rates itself.

To extend the analysis, the multivariate model was reestimated after including a money supply vector.

The inclusion of the money supply is suggested by the following reasoning. The observed correlations between the inflation and interest rate innovation may, to a certain extent, be explained by the joint correlations of both series with the money vector. The money supply series is the real monthly M1A series, transformed into a stationary form by the subsequent operations of a logarithmic transformation, first differencing, and seasonal differencing. The re-estimated multivariate model with a three-element $\hat{z}_t$ vector suggested the following form:

$$(I - \Phi_6 B^6)\hat{z}_t = (I - \theta_1 B^1 - \theta_2 B^2)(I - \theta_{12} B^{12})\hat{a}_t. \qquad (12)$$

The parameters in equation (12) were estimated using the conditional likelihood method. The estimated coefficients are presented in Table 1. Equation (12) indicates that there is a relatively strong correlation between the money supply and nominal interest rates. In fact, this correlation is much stronger than that between the interest and inflation rates. In general, all three series are driven by the innovation terms, but the lag structure is complex because of the multiplicative seasonal polynomial. The innovation lag structure is simplified by eliminating those parameters that are not significantly different from zero, but still the structure is complex.

This analysis indicates that the information set A also responds to money supply movements. The correlations among the innovations of the different series suggest that unsystematic information λ_i is simultaneously utilized in the bond, money, and commodity markets; and this information may be relevant in the determination of short-run decisions in all three markets. Money supply announcements seem to be a major factor in determining the short-run variations of nominal interest rates.

Conclusions

An alternative approach based on the previous work of Debreu and Hirshleifer is proposed for the analysis of nominal interest rates. The essential point of the analysis is the importance of new signals as a basis for decision making in an uncertain world. The analysis suggests that the Fisher hypothesis alone can be a too restrictive approach in explaining the behavior of nominal interest rates over time. The unsystematic information that changes the subjective probabilities assigned to the possible future states of the world can be at least as important as the information in the vector of past inflation rates. This suggests the possibility of an unstable real interest rate, even in the short-run. The test of this proposition versus the Fisher hypothesis, however, poses difficult problems. Part of these problems can be overcome by using recently developed multivariate stochastic analysis techniques. The application of these techniques to a monthly data set gave empirical results supporting this hypothesis. The results cannot be interpreted as conclusive given the subjective criteria involved in the testing procedure, but they are strongly suggestive of a mechanism more complex than the one implied by the Fisher hypothesis. This follows from the importance of information other than the actual inflation rate in the determination of nominal interest rates in a world of uncertainty.

TABLE 1

Estimation of the Extended Multivariate Model

Series 1: Actual Inflation Rate
Series 2: Nominal Interest Rates
Series 3: Real Money Supply (M1/P)
Method of Estimation: Conditional Likelihood

Φ_6

-0.414	0.001	-0.004
(0.058)	(0.001)	(0.007)
-1.716	-0.058	-0.019
(1.148)	(0.054)	(0.445)
0.008	0.015	0.249
(0.082)	(0.004)	(0.054)

θ_1

0.614	-0.006	-0.018
(0.060)	(0.002)	(0.016)
-6.780	-0.359	-2.572
(1.658)	(0.061)	(0.547)
0.044	0.005	-0.309
(0.109)	(0.005)	(0.058)

θ_2

0.081	-0.004	-0.015
(0.059)	(0.002)	(0.016)
-0.062	-0.002	-1.933
(1.738)	(0.063)	(0.481)
0.054	0.023	-0.219
(0.110)	(0.005)	(0.055)

θ_{12}

-0.148	0.006	-0.024
(0.064)	(0.002)	(0.021)
-0.590	-0.034	0.319
(1.867)	(0.065)	(0.610)
0.287	0.002	0.715
(0.134)	(0.005)	(0.046)

References

(1) Debreu, G., Theory of Value (New York: John Wiley and Sons, 1959).

(2) Fisher, I., The Purchasing Power of Money (New York: Macmillan, 1923).

(3) Fisher, I., The Theory of Interest (New York: Macmillan Company, 1930).

(4) Hirshleifer J., Investment, Interest and Capital, (New Jersey: Prentice-Hall Inc., 1970).

(5) Quenoille, M.H., Analysis of Multiple Time Series (New York: Hafner, 1957).

(6) Tiao, G.C. and G.E.P. Box, "Modeling Multiple Time Series with Applications", Journal of the American Statistical Association, 76, no: 376 (December, 1981), pp. 802-16.

RELATIVE FORECASTING PERFORMANCE OF AN ECONOMETRIC MODEL AND MULTIVARIATE TIME SERIES MODELS

by Asraul Hoque

Dept. of Econometrics,Monash Uni,Melbourne,Australia and

Dept. of Economics,Bilkent Uni,Ankara,Turkey

Abstract:The main focus of the study is to model and forecast four Australian quarterly macroeconomic series' Monetary Aggregate M1 , Real Gross Domestic Product (GDP) , Inflation rate and 90-day Bank Accepted Bill Rates (BAB). Forecasts of these variables are generated by a simple macroeconometric model,univariate time series model , multivariate VAR and BVAR models. To compare like with like all these variables are expressed in logs except the BAB rates. The data used to estimate these models are quarterly observations from 1970II to 1982IV. The out of sample forecasts are generated for the period 1983I - 1987III. The forecasting performance is judged by the Root Mean Squared Percentage Error (RMSPE).

In specifying the macromodel we have introduced Rational Expectations Hypothesis and an Error Correction technique to take into account of the long run economic theory in the dynamic context. Our simple model of the Australian economy consists of

$$\Delta m_t = \gamma_1 + \gamma_2 \Delta m_{t-1} + \gamma_3 R_{t-1} + \gamma_4 (m-p-y)_{t-1} + \epsilon_{1t}$$

$$y_t = \gamma_1 + \gamma_2 y_{t-1} + \gamma_3 R_{t-4} + \gamma_4 I_t + \gamma_5 (\Delta m_{t-4} - \Delta m^e_{t-4}) + \epsilon_{2t}$$

$$\Delta p_t = \gamma_1 + \gamma_2 \Delta p_{t-4} + \gamma_3 \Delta m_{t-4} + \gamma_4 (p+y-m)_{t-4} + \epsilon_{3t}$$

$$\Delta R_t = \delta_1 + \delta_2 \Delta^2 p_{t-2} + \delta_3 (R-p)_{t-2} + \epsilon_{4t}$$

where Δm_t = growth rate of money supply ; y_t = real GDP ;

Δp_t = inflation rate ; R_t = 90-day BAB rates .

The above specification has been retained after a good deal of experiments with different specifications. Now, in modelling the multivariate time series frameworks we use the same four variables expressed in logs except R. In this context we will consider three different VAR models using : (i) variables in log levels ; (ii) log levels plus trend components and (iii) difference-logged variables.The four VAR(p)models which have beenfound to be

adequateare VAR1(3) , VAR1(4) , VAR2(3) and VAR2(4). The first two refers to log levels with trend while the last two refers to difference logs. Overall , we choose VAR1(3) as the best adequate model for forecasting based on AIC , SC and HQ criteria. The results are also consistent with Sims LR test.

BVAR models have been chosen based on different combinations of prior parameters and adequacy test statistics of Box-Pierce and Ljung-Box. Finally , the Box-Jenkins approach is used to find a parsimonious ARIMA model for each of the time series. This could be used as a benchmark for comparison with other models. The models selected as most appropriate are ARIMA (1,1,1) , (1,1,0) , (1,1,1) and (2,1,1) for the series GDP , M1 , CPI , and BAB respectively .

Let us now briefly compare the forecasting performance. The out of sample forecasts of the macro model and VAR1 seem to predict the future path of actual real GDP quite well . The ARIMA model and to a lesser extent BVAR underpredict the actual values as the out of sample forecast horizon increases . For M1 , the out of sample forecast of the macro model is not significantly different from the actual values . BVAR stays quite close to the actual except for the 1987II - 1987III period . As the out of sample forecast horizon increases , VAR1 overpredicts the actual whereas ARIMA underpredicts M1 . For CPI the forecasting performance is quite close in the cases of VAR , BVAR and macro model . ARIMA overpredicts the actual for CPI. In the case of BAB , macro model does the best . The ARIMA , VAR and BVAR models miss most of the turning points by a significant amount for BAB.

The macro model produces better forecast than the ARIMA , VAR and BVAR for 92% , 100% and 58% of the cases respectively. The BVAR does better than unrestricted VAR for 75% of the cases. When we compare the univariate with the multivariate time series models , both VAR and BVAR perform better than the univariate ARIMA for 67% and 100% of the cases respectively .

SPECTRAL ANALYSIS OF TURKISH WHOLESALE PRICES

Andrzej Sokolowski

Bilkent University, Ankara, Turkey
Academy of Economics, Krakow, Poland

Turkish Wholesale Price Index covers 1422 commodities in 616 item groups. The whole market has been divided into 23 subsectors. Prime Ministry State Institute of Statistics publishes monthly values of wholesale price index for each subsector with base value 100 assigned to the average level of the year 1981. Data taken for research presented in this paper covers the period from January 1982 through August 1990 which gives 23 time series, each one consisting of 104 observations. Wholesale prices rose during analysed period (for half of subsectors more than 30 times) and all time series showed strong exponential trend. This trend component has been filtered by taking first difference of log of observations (resulting series is denoted by $\{x^1_{it}\}$, $(i=1,2,\ldots,23;\ t=1,2,\ldots,103)$). Primary hypothesis of research stated that all time series (subsectors) are generated by the same stochastic process. In order to verify this statement spectrum has been estimated for each time series. Parzen window was used with cut-off point m=24. Then distance matrix D between 23 time series has been calculated. Distance measure was defined as

$$d^2_{ij} = \frac{h}{\pi}\sum_{k=0}^{m}\left[\log\left[f^1_i(\pi k/m)/f^1_j(\pi k/m)\right]\right]^2$$

If spectral estimates $f^1_i(\omega)$ and $f^1_j(\omega)$ were calculated from series generated by the same process then d^2_{ij} follows chi-squared distribution with (m+1) degrees of freedom (Pinhas, Riviat [2]). Distance matrix D has been transformed into binary matrix B in which $b_{ij}=0$ when d^2_{ij} was greater than corresponding critical value taken from chi-squared distribution for significance level 0.05. Then the set of 23 subsectors has been clustered by means of vector eliminating method (Chomatowski, Sokolowski [1]). This method creates groups with elements (objects) which are mutually non-different. In other words the smallest number of all-zero submatrices are taken out of matrix B with all main diagonal elements to be selected. We have found that Turkish wholesale market could be divided into three segments, so there are three different stochastic processes which generate the behaviour of wholesale price index.

Segment I

There are only fishery products in that segment. Primarily transformed data $\{x_{it}\}_{(1)}$ can be described by first-order autoregression process with additional periodic component of 3 month.

Segment II

This segment consists of two subsectors, textile and metal machinery & vehicles. The spectrum of mean values (from two time series) showed no seasonality but strong trend has been revealed. First differences of $\{x^*_t\}_{(2)}$ are well fitted by first-order moving average process. Variance of $\{x^1_t\}_{(2)}$ is stationary.

Segment III

This is the main body of Turkish wholesale market, with 20 subsegments inside, so there are 20 time series as the realizations of the same stochastic process. Autocovariance matrix has been estimated. Elements at main diagonal possessed significant positive trend, what meant that variance within this segment was not stationary. The analysis of subdiagonals (up to the lag 24) proved that there was no trend in covariances. Then autocovariance function has been estimated as the average and mode from the respected subdiagonals of autocovariance matrix. Both estimators led to the same conclusions. The estimated spectrum showed important seasonal variations. Because for each time point we had 20 observations, then we were able to study the time series of sample characteristics of marginal distributions of analysed stochastic process. The best model for mean value was first order autoregression with seasonal component. The behaviour of the mode was described by just seasonal autoregression $SAR(12)_1$. The asymetry measure showed slightly declining trend.

References

[1] Chomatowski S., Sokolowski A., Taksonomia struktur, *Przegląd Statystyczny*, 1978, Nr 2

[2] Pinhas G.S., Riviat P.J., The Analysis of Simulation-Generated Time Series, *Management Science*, 1981, No 1

COINTEGRATION AND ERROR CORRECTION IN COMMON STOCK PRICES: THE CASE OF ISTANBUL STOCK EXCHANGE

HUSEYIN KELEZOGLU
Bilkent University
P.O.B. 8, Maltepe.
Ankara, TURKEY

AND

ERDEN BASCI
Bilkent University
P.O.B. 8, Maltepe
Ankara, TURKEY

ABSTRACT-Security prices cannot be cointegrated if they are generated in an efficient market. The existence of an error correction representation of security prices would imply the improvement of the return forecasts which is against the weak form efficiency. In this study, we investigate the dynamic behaviour of prices of three industry based portfolios of common stocks traded in the Istanbul Stock Exchange (ISE) for the 1988-89 period. Although no pairwise cointegration is detected, we reject the nonstationarity of a linear combination of the three prices. In addition, the estimation of the error correction representation of these prices is carried out. A trading rule based on this representation gave profits which are higher than obtainable from the "buy and hold" strategy which increases our doubts about the weak-form efficiency of the ISE.

I. INTRODUCTION

One of the famous and empirically regular instances of random walk occurs in stock price series. In the light of the recent developments in the theory of cointegration and error correction mechanisms, led by Granger and Weiss(1983) Engle and Granger (1987) Johansen(1988), we know that certain linear combinations of more than one nonstationary time series, such as random walks, may be stationary. Such nonstationary series are said to be cointegrated. Granger and Engle (1987) suggest estimating the cointegrating vector by OLS regressions. The properties of the estimates of this cointegrating regression have been investigated by Stock(1987), Phillips(1987) and a number of other authors. If there are cointegrated stock prices, there would be an error correction representation of at least one of them, which could be used for forecasting purposes. In an efficient market, as is well known, such forecasts are of no value.

In this study, the main aim is an empirical investigation of cointegration and error correction(EC) in the context of stock prices rather than driving economic conditions under which stock prices are cointegrated.

The Data: The data used in our investigation is the weekly prices of three industry based common stock portfolios traded in Istanbul Stock Exchange for the period 1988-89 (104 observations). The particular stocks are chosen such that the data is continuously available for the period under consideration. The industries are Chemical Industry (CI) which consists of five stocks: Baspetr, Eggubre, Kav, Gubre Fabrikalari, Koruma Turle, Finance and Banking Industry (FB) which comprise four stocks, Eczacibasi, Koc Yatirim, Koc Holding, T Isbankasi(b); Metal Industry (MI) which consists of six stocks Celikhalat, Erdemir, Nasas, Rabak, Sarkuysan and Demirdokum.

II. COINTEGRATION IN STOCK PRICES

In accordance with our expectations, the hypothesis of random walk is not rejected for any of the price series. The test statistics we used are Dickey-Fuller(DF) and Augmented Dickey-Fuller(ADF). For the ADF test, lag lengths are taken up to four and a general to specific approach is followed.

For the super-consistency result of Stock(1986) to apply, no subset of the explanatory variables should be cointegrated. So, we have first investigated whether there are any mutually cointegrated stock prices or not. Results are presented in Table I. Regressing each of the prices on another gave Cointegrating Regression Durbin Watson (CRDW). DF, ADF statistics which are well below their critical values even at a 10% significance level[1]. These lead us to conclude that there exists no pairwise cointegration between these three portfolio prices.

To detect any cointegrating vector which binds the three prices we have run three cointegrating regression every time making one of the prices the dependent variable. The results are reported in Table.II. Looking at the results, we see that in all cases we reject the null of non-cointegration.

TABLE I Results of pairwise cointegration regressions

Dep Vars	CI	FB	MI
CI	-	1.56(27.26)	
FB	-	-	1.06(18.56)
MI	0.49 (18.52)	-	-
Constant	4.33 (22.24)	0.88(5.80)	-0.58(-3.22)
R^2	0.79	0.96	0.95
CRDW	0.17	0.20	0.20
DF	-1.99	-2.52	-2.78
ADF	-2.53	-1.99	-2.18
σ	0.13	0.14	0.18

NOTE: t-values are in parentheses. All variables are in natural logs.

We use the critical values of ADF and CRDW tests for the three variable case provided by Hall(1986), and of DF test for the two variable case provided by Engle and Granger (1987). Compared to the critical values of these tests, the test statistics are highly significant. For example, the regression of FB on CI & MI gave CRDW, DF, ADF statistics which are significant even at 1% level of significance. Furthermore, all the regression parameters are highly significant supporting the result of cointegration. The cointegration relation seems to be the highest when the dependent variable is FB since R^2, CRDW, DF are highest in all cases compared to that of the other two. Stock(1986) establishes that the estimates of the cointegrating regression are consistent and subject to a smaller bias than the OLS regression of stationary series, if the series are cointegrated. Furthermore, this bias seems to be related to the overall goodness of fit of the regression, and so the cointegrating regression with the highest R^2 is expected to be subject to the smallest bias. We notice that in Table II the regression of FB on CI & MI gave DF statistics which is the highest compared to that of the other two.

[1] The critical values at a 10% level of significance are CRDW= 0.22, DF=-2.57, ADF=-2.58 from Engle and Granger(1987).

TABLE II Results of three variable cointegrating regressions

Dep.Vars	CI	MI	FB
CI	-	-0.427 (-4.19)	0.56 (10.32)
MI	-0.346 (-4.18)	-	0.63 (21.36)
FB	0.922 (10.32)	1.294 (21.37)	-
Constant	3.532 (22.35)	1.015 (2.439)	-1.54 (-5.98)
R^2	0.90	0.96	0.96
CRDW	0.46	0.35	0.49
DF	-3.63	-3.5	-4.01
ADF	-4.19	-4.2	-4.00
σ	0.12	0.13	0.09

NOTE: t-values are in parentheses. All variables are in natural logs.

ERROR CORRECTION REPRESENTATION OF STOCK PRICES

Although the test results show that the most suitable cointegration relation occurs when FB is the dependent variable we form an error correction representation for all the three cointegrating vectors above so that we see clearly the relation between an error correction and cointegration mechanism. The results are presented in Table III. We proceed by applying a general to specific approach and reporting only the significant lags and variables.

TABLE.III Results of error correction estimation

Dep.Vars.	ΔCI	ΔFB	ΔMI
ECI(-1)	-0.19 (2.92)	-	-
EFB(-1)	-	-0.11 (-1.64)	-
EMI(-1)	-	-	-0.02 (-0.45)
F(-1)	-	0.41 (4.47)	-
F(-3)	0.59 (3.86)	0.40 (3.84)	0.53 (5.52)
M(-3)	-0.30 (-1.9)	-	-
M(-4)	-0.32 (-2.1)	-0.28 (-2.45)	-0.28 (-2.48)
Constant	0.01 (1.37)	-	0.02 (2.63)
R^2	0.22	0.27	0.23
DW	1.64	1.93	1.44
σ	0.08	0.08	0.08

NOTE: t-values are in parentheses. All variables are in natural logs. The variables ECI, EMI and ECI represent the residuals derived from the cointegrating regressions of FB, CI and MI reported on Table.II. We proceed here by applying a general to specific approach.

The results in the above table clearly suggest that the error correction representation is supported by the series. The error correction terms are significant in two of the regressions, namely ΔFB and ΔCI. The insignificance of the error term in the ΔMI regression implies the weak exogeneity of the metal industry portfolio prices in the sense of Engle and Granger (1987).

To be able to question weak form pricing efficiency however these R-squared figures are not sufficient. The value of a forecast is a function of how it is used, as well as how precise it is. So one has to explain the decision rule, in our case a trading rule which employs the forecasts generated by the error correction representation given above. We use here the following simple strategy. In case the expected change in the stock's price is positive, hold 100% of your wealth in that stock, otherwise hold 100% money which is assumed to have zero return. Transaction costs are assumed to be negligible. This trading rule is simulated on our data for the same period. We simulate the wealths of three different investors who start with all their wealth invested in one unit of one of the sector portfolios. Each of them is assumed to follow the same trading rule based on the error correction representation of the sector in question. The results are summarized in Table IV. All of our three investors have "beaten the market" in the sense that their strategy has provided higher returns than the buy and hold strategy for the two years. The excess return is the greatest (31.2%) for the chemistry industry. The metal industry which was found to be weakly exogeneous, gave the lowest excess return (5.3%).

Table IV Results of trading rule simulations

	Starting Wealth	Ending Wealth: Buy and Hold	Ending Wealth: Trade	Excess Return (%)
Chemistry	8.389	9.421	9.733	31.2
Metal	7.672	9.642	9.695	5.3
F.& Banking	8.170	9.922	10.087	16.0

NOTE: All wealths are in natural logs.

These findings indicate the possibility of cointegrated common stock prices. The rejection of the weak form pricing efficiency however does not readily follow. Our evidence from the trading rules however increases our doubts on the pricing efficiency of the Istanbul Stock Exchange.

REFERENCES

Engle R.F. and Granger, C.W.J (1987) "Co-integration and error correction: Representation, estimation and testing," Econometrica 55, 251-276.

Granger C.W.J. and Weiss,A.A (1983) "Time series analysis of error correcting models," in Studies in Econometrics, Time Series and Multivariate Statistics, New York, Academic Press, UCSD discussion paper, 255-278.

Hall, S.G (1986) "An application of the Granger and Engle two step estimation procedure to United Kingdom aggregate wage data," Oxford Bulletin of Economics and Statistics, 48, 220-239.

Johansen,S. (1988) "Statistical analysis of cointegration vectors," Journal of Economic Dynamics and Control, 12, 231-254, North Holland.

Phillips, P.C.B. (1987) "Time series regression with a unit root," Econometrica, 55, 277-301.

Stock, J.H. (1987)"Asymptotic properties of least squares estimators of cointegrating vectors," Econometrica 55, 1035-1056.

Dynamic Oligopolies as Neural Networks

by

Ferenc Szidarovszky

Abstract

Multiproduct oligopoly models are investigated under the additional assumptions that if the difference between the present and profit maximizing outputs is too small, then no output exchange is performed, and if the same difference is too large, then only a portion of the desired output change is performed by the firms. These binary variables transform the classical oligopoly models into special neural networks. The stability of these dynamic systems is analyzed and a least squares learning procedure is outlined.

1. Introduction

The theory of oligopoly has been investigated very intensively by researchers in mathematical economy. The existence and uniqueness of the equilibrium in oligopoly markets were discussed, for example, by Szidarovszky and Yakowitz (1977, 1982). A survey on the different variants of oligopoly models and their properties is given in Okuguchi and Szidarovszky (1990), with a comprehensive literature summary.

In this paper a new version of single-product oligopolies is introduced, where the mathematical model can be interpreted as a special neural network. A solution methodology is suggested to discuss the stability of the resulting dynamic system. We note that a special case of this model was investigated earlier by Szidarovszky and Okuguchi (1988).

2. The Mathematical Model

Let n denote the number of firms without product differentiation in an oligopoly. Let $p = a - bQ$ ($a>0$, $b>0$) be the market demand function, where $Q = \sum_{i=1}^{n} x_i$ is the sum of all firm's output, x_i is the i-th firm's output and p is the common market price of the goods produced by the firms. We assume that the market demand function is completely known to all firms, but that they do not have complete information on the other firms' marginal cost functions. They are therefore obliged to perceive other firms' marginal cost functions which are consistent with the past information on the firm's outputs. Let the j-th firm's marginal cost perceived by the i-th firm be linear and given by $MC_j^i = +c_j^i + m_j^i x_j$, $i \neq j$, $j = 1,2,...,n$, where c_j^i and m_j^i are both nonnegative constants. Assuming that the i-th firm thinks that at any period the j-th firm selects its profit maximizing output, which is assumed to be interior, the marginal revenue and marginal cost are equal:

$$a - b(x_j + Q_j) - bx_j = c_j^i + m_j^i x_j \quad (i \neq j), \tag{1}$$

where $Q_j = \sum_{k \neq i,j} x_k$ is the output of the rest of the industry. Let y_j^i be the i-th firm's expectation of the j-th firm's output. At time t when the i-th firm forms expectations about the other firm's outputs on the basis of its perceived marginal cost functions, the following similar equation has to hold:

$$a - b\left(x_i(t) + \sum_{e \neq i,j} y_e^i(t)\right) - by_j^i(t) = c_j^i + m_j^i y_j^i(t), \quad (i \neq j) \tag{2}$$

On the other hand, at time t, outputs of all firms at the previous time $t-1$ are known, and they must be consistent with (2). Hence

$$a - b\left(x_i(t-1) + \sum_{e \neq i,j} x_e(t-1)\right) - bx_j(t-1) = c_j^i + m_j^i x_j(t-1) \quad (i \neq j). \tag{3}$$

By solving these equations for $y_j^i(t)$ simple calculation shows that the expected profit maximizing output $x_i(t)$ of firm i can be expressed as

$$x_i(t) = g_i(x_i(t-1), Q_i(t-1)), \quad i = 1,2,...,n. \tag{4}$$

This is the reaction function of the i-th firm derived on the basis of its perceived marginal cost functions of all other firms which are consistent with the information on all firms' outputs from the preceding time period.

3. The Modified Model

Assume next that each firm specifies a positive number δ_i, which has the following meaning. If the profit maximizing output (4) does not differ from the output $x_i(t-1)$ of the preceding time period more than δ_i, then firm i does not change its output. It is also assumed, that each firm specifies another positive constant $\Delta_i > \delta_i$, which is the maximal allowed change in its output during a single time period. The first assumption can be interpreted as very small changes are not made, since the costs of changing the output at all are higher than the loss by selecting suboptimal outputs. The second assumption means that the changes in the output during a single time period are bounded. Based on these additional assumptions the following dynamic process is obtained:

Let $x_1(0), x_2(0), ..., x_n(0)$ denote the initial outputs of the firms. For $t \geq 1$, define $x_i^*(t) = g_i(x_i(t-1), Q_i(t-1))$. Then the output of the i-th firm at time period t is given as

$$x_i(t) = \begin{cases} x_i(t-1), & \text{if} \quad x_i(t-1) - \delta_i \leq x_i^*(t) \leq x_i(t-1) + \delta_i \\ x_i^*(t) & \text{if either} \quad x_i(t-1) - \Delta_i \leq x_i^*(t) < x_i(t-1) - \delta_i \\ & \text{or} \quad x_i(t-1) + \delta_i < x_i^*(t) \leq x_i(t-1) + \Delta_i \\ x_i(t-1) - \Delta_i, & \text{if} \quad x_i^*(t) < x_i(t-1) - \Delta_i \\ x_i(t-1) + \Delta_i, & \text{if} \quad x_i^*(t) > x_i(t-1) + \Delta_i \end{cases} \tag{5}$$

The global asymptotic stability of this dynamic system can be discussed in a similar manner as it is documented in Liu and Szidarovszky (1991), and by using the theory of perturbed contractions (Szidarovszky, and Yakowitz, 1978).

This model can be easily accompanied with a special learning process on the unknown parameters c_j^i and m_j^i. Assume first that at each time period t, each firm measures the outputs of all other firms at all preceding time periods. Then from equation (3) each firm has the time series $V_j^i(t-r) = c_j^i + m_j^i x_j(t-r)$ for $r = 1,2,...,t$, from which the values of c_j^i and m_j^i can be obtained, for example, by using a least squares method. Since at each new time period a new value $V_j^i(t) = c_j^i + m_j^i x_i(t)$ is computed from equation (3), the parameter values can be updated, and these updated values can be used in forming the new expectations. If the individual outputs can not be measured, only the $Q_i(t-r)$ values, $r = 1,2,...,t$, then the following procedure can be proposed. By using equation (3) and the definition of $Q_i(t)$ an overdetermined nonlinear equation system is obtained:

$$c_j^i + m_j^i x_j(t-r) = V_j^i(t-r),\ i \neq j$$
$$\sum_{j=1, j\neq i}^{n} x_j(t-r) = Q_i(t-r),\ r = 1,2,...,t, \qquad (6)$$

The solution of this overdetermined system can be obtained by, for example, a nonlinear least squares method.

4. Conclusions

The above dynamic model is a typical example for a neural network. The two major features of this model are the adaptive learning procedure and the presence of the binary variables determining which case is to be used in obtaining the new values of $x_i(t)$ in relation (5).

The dynamic system discussed in this paper is a generalized version of earlier works in oligopoly theory.

References

Grossberg, S. (1988) Nonlinear Neural Networks: Principles, Mechanisms, and Architectures. Neural Networks, Vol. 1, pp. 17-61.

Liu, D. and F. Szidarovszky (1991) Global Asymptotic Stability of Dynamic Systems with Modified Contractions. Appl. Math. and Comp. (to appear).

Okuguchi, K. and F. Szidarovszky (1990) The Theory of Oligopoly with Multi-Product Firms, Springer-Verlag, Berlin/Heidelberg/New York.

Szidarovszky, F. and K. Okuguchi (1988) Perceived Marginal Costs in an Oligopoly Model. Paper presented at the 3rd Congress of the European Economic Association, Bologna, Aug. 27-29, 1988.

Szidarovszky, F. and S. Yakowitz (1977) A New Proof of the Existence and Uniqueness of the Cournot Equilibrium. Intern. Econ. Review, Vol. 18, pp. 787-789.

Szidarovszky, F. and S. Yakowitz (1978) Principles and Procedures of Numerical Analysis. Plenum Press, New York/London.

Szidarovszky, F. and S. Yakowitz (1982) Contributions to Cournot Oligopoly Theory. J. of Econ. Theory, Vol. 28, pp. 51-70.

THE POSSIBILITY OF OUTPUT FLUCTUATIONS IN DUOPOLY MODELS

PROFESSOR CARL CHIARELLA
SCHOOL OF FINANCE AND ECONOMICS
UNIVERSITY OF TECHNOLOGY, SYDNEY

1. Introduction

The literature on the stability of the Cournot model, in both discrete and continuous time, is one of the richest in economic dynamics. See, eg. Okuguchi and Szidarovsky (1990). In recent years the advances in the qualitative theory of differential equations and the theory of nonlinear dynamical systems (see eg. Arnold (1978) and Guckenheimer and Holmes (1983)) have led to a consideration of the dynamic behaviour of the Cournot model when the equilibrium point is locally unstable. The key studies in this vein are Seade (1980), Al-Nowaihi and Levine (1985) and Furth (1986). In the cited works, quite general Cournot oligopoly models were considered and some quite general qualitative results about the dynamic behaviour were obtained.

Our aim in this paper is more modest. We consider a duopoly model in its region of local instability and explore the possibility that output tends to limit cycle motion. Our discussion here is necessarily brief, however, a more detailed examination of limit cycle motion in duopoly models is contained in Chiarella (1991).

2. Some Possible Nonlinear Mechanisms

To obtain sustained output fluctuations in duopoly models we will need to introduce time lags and nonlinearities into the standard formula. *Time lags* can be introduced as in Chapter 7 of Okuguchi (1976) where firms adjust their output to desired output with a lag and use an adaptive expectations scheme to form their expectation of the other firms output. A range of *nonlinearities* is possible and include:- (a) introducing cubic and higher order terms into the cost function, and/or, introducing quadratic and higher order forms into the demand function; (b) introducing constraints on output adjustment. Suppose for example that $\bar{x}_i$ is the desired output of firm i and that the firm adjusts to the desired level with a lag according to: $\dot{x}_i = k_i (\bar{x}_i - x_i)$, $k_i>0$. However, it may be costly for the firm to make large output changes, indeed output changes above or below certain levels may be impossible. This effect could be captured by assuming a relationship between $\dot{x}_i$ and x_i as shown in Figure 1.

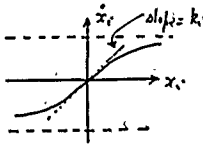


Figure 1

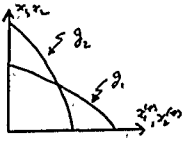


Figure 2

(c) An alternative way to capture the effect in (b) would be to include an $\dot{x}_i^2$ term in the cost function, eg. $C_i(x_i, \dot{x}_i) = C_i x_i + d_i x_i^2/2 + e_i \dot{x}_i^2/3$, $(C_i>0, d_i \geq 0, e_i \geq 0)$.

3. Simple Lag Structure

In order to perform an initial analysis, let's assume a nonlinearity of the type (a) above. Suppose demand is given by

$$p = a - b(x_1 + x_2), \quad (b>0) \tag{1}$$

and the cost function for firm i is given by

$$C_i(x_i) = c_i x_i + d_i x_i^2/2 + e_i x_i^3/3, \quad (c_i \geq 0, d_i>0, e_i \geq 0) \tag{2}$$

The profit function for firm 1 is:

$$\Pi(x_1, x_2^{(e)}) = [a - b x_1 + x_2^{(e)})]\, x_1 - C_1(x_1) \tag{3}$$

where $x_2^{(e)}$ is firm 1's expectation of firm 2's output.

The profit maximizing condition $\frac{\partial \Pi_1}{\partial x_1} = 0$ implies:

$$(2b + d_1) x_1 + e_1 x_1^2 = (a - c_1) - b x_2^{(e)}, \tag{4}$$

from which

$$x_1 = g_1[x_2^{(e)}] \tag{5}$$

where g_1 is a downward sloping reaction function which is displayed in Figure 2. Similarly from the profit maximising condition of firm 2, we find that:

$$(2b + d_2) x_2 + e_2 x_2^2 = [a - C_2(x_2)] - b x_1^{(e)}, \tag{6}$$

ie. $x_2 = g_2[x_1^{(e)}]$ (7)

This reaction function is also illustrated in Figure 2. We assume that each firm adjusts to the desired level of output with a lag, ie.

$$\dot{x}_1 = k_1 [g_1(x_2^{(e)}) - x_1], \tag{8a}$$

$$\dot{x}_2 = k_2 [g_2(x_1^{(e)}) - x_2]. \tag{8b}$$

where the k_i (>0) $(i = 1,2)$ are speed of adjustment parameters.

We assume that each firm forms its expectation of the other firm's output adaptively. Thus, for firm 1 we would write:

$$x_2^{(e)}(t) = \frac{1}{T_2} \int_0^t e^{-(t-s)/T_2} \, x_2(s)\, ds, \tag{9}$$

where $T_2 > 0$ may be interpreted as a time lag in the formation of expectations. Equation (9) may be expressed in differential form:

$$\dot{x}_2^{(e)} = (x_2 - x_2^{(e)})/T_2. \tag{10}$$

The dynamics of the economy are then governed by the four-dimensional differential system:

$$\dot{x}_1 = k_1 [g_1(x_2^{(e)}) - x_1], \quad \dot{x}_2 = k_2 [g_2(x_1^{(e)}) - x_2], \tag{11a}$$

$$\dot{x}_1^{(e)} = \frac{1}{T_1}[x_1 - x_1^{(e)}], \quad \dot{x}_2^{(e)} = \frac{1}{T_2}[x_2 - x_2^{(e)}]. \tag{11b}$$

4. Analysis of the Dynamics

The equilibrium point of the differential system (11) is given by:

$$\bar{x}_1 = g_1(\bar{x}_2), \quad \bar{x}_2 = g_2(\bar{x}_1), \tag{12a}$$

$$\bar{x}_1^{(e)} = \bar{x}_1, \quad \bar{x}_2^{(e)} = \bar{x}_2. \tag{12b}$$

The *local* stability properties of the model are determined from the Jacobian of the differential system (2), which is given by:

$$J_1 = \begin{bmatrix} -k_1 & 0 & 0 & k_1 g_1' \\ 0 & -k_2 & k_2 g_2' & 0 \\ \frac{1}{T_1} & 0 & -\frac{1}{T_1} & 0 \\ 0 & \frac{1}{T_2} & 0 & -\frac{1}{T_2} \end{bmatrix} \tag{13}$$

To get a feeling for how the four dimensional system is behaving, consider the special case in which both firms are perfectly identical so that,

$$k_1 = k_2 = k, \; g_1 = g_2 = g. \tag{14}$$

Then we need only consider the two dimensional system.

$$\dot{x} = k[g(x^e) - x] \text{ and } \dot{x}^e = \frac{1}{T}(x - x^e), \tag{15}$$

which has the simpler Jacobian matrix

$$J_2 = \begin{bmatrix} -k & kg' \\ \frac{1}{T} & -\frac{1}{T} \end{bmatrix} \tag{16}$$

We see that:

$$|J| = \frac{k}{T}(1 - g') > 0, \text{ and } \mathrm{tr}(J) = -(k + \frac{1}{T}) < 0. \tag{17}$$

So the equilibrium is stable for all $T \geq 0$. Hence, no output fluctuations will occur in the model when the two firms become identical. In this preliminary study, we do not consider the stability properties when the two firms are allowed to differ.

5. Higher Order Lag Structure

Now let's introduce the higher order lag structure:

$$x^{(m)}_{(t)} = \int_0^t \frac{e^{-(t-s)/T}}{T} [x(s) - x^{(m)}(s)]\, ds, \tag{18}$$

The expectations mechanism in (18) (with $m > 0$) takes past differences between $x^{(m)}(s)$ and $x(s)$ ($s \leq t$) and weights these with exponentially declining weights from t to 0. This scheme is a way of capturing approximately in continuous time, the lagged expectations scheme:

$$\dot{x}(t) = m\,[x(t-T) - x^{(m)}(t-T)].$$

It can be shown that in the limit $T \to 0$, equation (18) yields the standard adaptive expectations scheme:

$$\dot{x}^{(m)}(t) = m\,[x(t) - x^{(m)}(t)].$$

In (19) set

$$y(t) = \int_0^t \frac{e^{-(t-s)/T}}{T} [x(s) - x^{(m)}(s)]\, ds, \tag{19}$$

which satisfies

$$Ty(t) = [x(t) - x^{(m)}(t)] - y(t). \tag{20}$$

We then replace (19) with (21) and

$$\dot{x}^{(m)}(t) = my(t). \tag{21}$$

If we consider the case of identical firms, we have the three dimensional system:

$$\dot{x} = k\,[g(x^e) - x], \tag{22a}$$

$$\dot{x}^e = my, \tag{22b}$$

$$\dot{y} = \frac{1}{T}(x(t) - x^e(t)) - \frac{1}{T}y(t). \tag{22c}$$

The Jacobian of this differential system is:

$$J_3 = \begin{pmatrix} -k & kg' & 0 \\ 0 & 0 & m \\ \frac{1}{T} & \frac{-1}{T} & \frac{-1}{T} \end{pmatrix} \tag{23}$$

The characteristic equation is:

$$\lambda^3 + \left(k + \frac{1}{T}\right)\lambda^2 + \frac{(m+k)}{T}\lambda + \frac{mk}{T}(g' - 1) = 0. \tag{24}$$

Note that the product of the roots is:

$$\lambda_1 \lambda_2 \lambda_3 = \frac{mk}{T}(g' - 1) < 0. \tag{25}$$

In order to apply Hopf bifurcation theory we need to establish the existence of a pair of pure imaginary roots, $\lambda = iw$. In such a case, w would have to satisfy:

$$-iw^3 - \left(k + \frac{1}{T}\right)w^2 + i\frac{(m+k)}{T}w - \frac{mk}{T}(g' - 1) = 0 \tag{26}$$

ie. we would require:

$$w^2 = \frac{m+k}{T} \quad \text{and} \quad w^2 = \frac{-mk}{kT+1}(1-g') \tag{27}$$

Equating the two expressions for w^2, we find the value of T at which pure imaginary roots can occur, viz.

$$T^* = \frac{-(m+K)}{k(mg' + k)}. \tag{28}$$

To have $T^* > 0$ we would require,

$$k < -mg'. \tag{29}$$

The conditions (29) poses an upper limit on the speed adjustment of output to desired output. Differentiating (25) with respect to T and setting $T = T^*$ reveals that

$$R\left(\frac{\partial \lambda}{\partial T}\right)_{T=T^*} = \frac{w^2}{2T^2[w^2 + (k + 1/T)^2]} > 0 \tag{30}$$

Thus by the Hopf bifurcation theorem (see eg. Guckenheimer and Holmes (1983)) we can assert the existence of a limit cycle for T in the neighbourhood of $T = T^*$. Since the $R(\lambda)$ increases as T increases through T^* the model displays the dynamic behaviour illustrated in Figure 3.

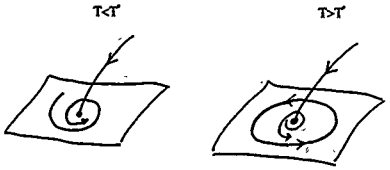


Figure 3

6. Conclusion

We have seen that by introducing nonlinearities into firms' cost functions and allowing for a sufficiently high order lag structure in the formation of expectations, it is possible for the duopoly model to exhibit fluctuating output in the form of a limit cycle.

Further research could consider the situation when the two firms are not identical and extend the analysis to oligopoly models.

REFERENCES

Al-Nowaihi, A. and Levine, PL (1985), *"The Stability of the Cournot Oligopoly Model: A Reassessment"*, Journal of Economic Theory, 35, 307-321

Arnold, V.I. (1978), *"Ordinary Differential Equations"* , MIT Press

Chiarella, C. (1991), *"The Instability of the Cournot Model"*, Working Paper, School of Finance and Economics, University of Technology, Sydney.

Furth, D. (1986), *"Stability and Instability in Oligopoly"*, Journal of Economic Theory, 40, 197-228

Guckenheimer, J. and Holmes, P. (1983), *"Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields"*, Springer-Verlag, New York.

Okuguchi, K. (1976), *"Expectations and Stability in Oligopoly Models"*, Lecture Notes in Economics and Mathematical Systems, Vol. 138, Springer-Verlag.

Okuguchi, K. and Szidarovsky, F. (1990) *"The Theory of Oligopoly with Multi-Product Firms"*, Lecture Notes in Economics and Mathematical Systems, Vol. 343, Springer-Verlag.

Seade, J. (1980), *"The Stability of Cournot Revisited"*, Journal of Economic Theory, 23, 15-27.

ON THE GLOBAL STABILITY OF THE COURNOT EQUILIBRIUM UNDER PRODUCT DIFFERENTIATION

KOJI OKUGUCHI
Department of Economics,
Tokyo Metropolitan University
1-1-1 Yakumo, Meguro-ku, Tokyo, JAPAN
KAZUYA IRIE
Institute of Social Sciences,
The University of Tsukuba
Tsukuba-shi, Ibaraki-ken, JAPAN

I. INTRODUCTION

Until quite recently, it has been assumed in most works on the global stability of the Cournot oligopoly equilibrium that the rates of change of actual outputs are sign-preserving functions of the divergences between the expected profit-maximizing and actual outputs. An alternative output adjustment system may be formulated, where the firms are assumed to increase, decrease or do not change their actual outputs depending on the signs of the marginal profits with respect to their own outputs.

In this paper we will show the stability-wise equivalence of the two dynamic output adjustment systems. We also introduce an alternative adjustment system where outputs will remain nonnegative, and prove the global stability under strict monotonicity of the marginal revenue functions.

II. COURNOT DUOPOLY

In this section we consider Cournot duopoly with product differentiation. Let $p_i=f^i(x_1,x_2)$ and $C_i(x_i)$ be the inverse demand and cost functions for the i-th firm, p_i and x_i be the price and output of the i-th firm, where i=1,2. The i-th firm's profit π_i is given by

(1) $\pi_i \equiv x_i f^i(x_i,x_j)-C_i(x_i)$, $i\neq j$, $i,j=1,2$.

Assumption 1. $\partial f^i/\partial x_i \equiv f^i_i < 0$, $f^i \in C^2$ for $(x_1,x_2) \in [0,\bar{x}_1)\times[0,\bar{x}_2)$, where $f^i(x_i,x_j)=0$ for $x_i \geq \bar{x}_i$ and for any x_j, $i\neq j$, $i,j=1,2$.

Assumption 2. The second order condition is satisfied for the two firms, that is,

(2) $\partial^2\pi^i/\partial x_i^2 \equiv \pi^i_{ii} = f^i+x_i f^i_i+f^i_i-C'_i < 0$, $i=1,2$.

The cross partial derivative, π^i_{ij}, $i\neq j$, however, may take any sign. If two goods are substitutes, $f^i_j<0$ but the sign of f^i_{ij} is indeterminate.

The firms' actual outputs are assumed to be adjusted over time according to (3a) or (3b):

(3a) $dx_i/dt = g_i(\partial\pi_i/\partial x_i)$, $g^i(0)=0$, $g^i_i>0$, $i=1,2$,

(3b) $dx_i/dt = h_i(y_i-x_i)$, $h_i(0)=0$, $h^i_i>0$, $i=1,2$,

where y_i defined by (5) satisfies

(4) $f^i(y_i+x_j)+y_i f^i_i(y_i+x_j)-C'_i(y_i) = 0$, $i\neq j$, $i=1,2$.

(5) $y_i \equiv \psi^i(x_j)$, $i\neq j$, $i,j=1,2$.

First, we derive for the adjustment process (3a):

(6·1) $dx_2/dx_1 = -\pi^1_{11}/\pi^1_{12}$ for $dx_1/dt=0$,

(6·2) $dx_2/dx_1 = -\pi^2_{21}/\pi^2_{22}$ for $dx_2/dt=0$.

In view of (2), (6·1) and (6·2), the following five cases are possible.

Case 1. $\pi^1_{12}<0$, $\pi^2_{21}<0$, $\pi^1_{11}\pi^2_{22}>\pi^1_{12}\pi^2_{21}$.

Case 2. $\pi^1_{12}<0$, $\pi^2_{21}<0$, $\pi^1_{11}\pi^2_{22}<\pi^1_{12}\pi^2_{21}$.

Case 3. $\pi^1_{12}>0$, $\pi^2_{21}>0$, $\pi^1_{11}\pi^2_{22}>\pi^1_{12}\pi^2_{21}$.

Case 4. $\pi^1_{12}>0$, $\pi^2_{21}>0$, $\pi^1_{11}\pi^2_{22}<\pi^1_{12}\pi^2_{21}$.

Case 5. $\pi^1_{12}>0$, $\pi^2_{21}>0$.

We can easily observe on the basis of the phase diagrams that the Cournot equilibrium is globally stable in Cases 1, 3 and 5, but unstable (saddle point) in Cases 2 and 4, and that no cycle arises. That there arises no cycle has been observed earlier by Furth (1986) for Cournot oligopoly without product differentiation.

We now consider the alternative adjustment process (3b). Taking into account $h^i_i>0$ and $\pi^i_{ii}<0$, we get

$$\operatorname{sgn} dh_i = \operatorname{sgn}(\pi^i_{ii}dx_i+\pi^i_{ij}dx_j), \quad i=1,2.$$

This shows the stability-wise equivalence of the two adjustment systems (3a) and (3b).

III. COURNOT OLIGOPOLY WITH BOUNDARY CONDITION

Let there be n firms producing differentiated goods, and let $p_i \equiv f^i(x_1,\cdots,x_n)$ and $C_i(x_i)$ be the i-th firm's price and cost functions, where x_i is the i-th firm's output. A finite positive number $\bar{x}_i$ has a similar property as in the preceding duopoly. Let Assumptions 1-3 be valid also for oligopoly, where π_i is now defined as

(7) $\pi_i = x_i f^i(x_i,x_{-i})-C_i(x_i)$, $i=1,2,\cdots,n$,

where $x_{-i} = (x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)^T$.

Assumption 4. There exists a unique interior Cournot equilibrium.

The boundary condition is stated as

Assumption 5. $\partial\pi_i/\partial x_i > 0$, for $x_i=0$ and any x_{-i}, and $\partial\pi_i/\partial x_i < 0$, for $x_i=\bar{x}_i$ and any x_{-i}, $i=1,2,\cdots,n$.

As in Section 2, we consider two alternative output adjustment systems for $i=1,2,\cdots,n$

(8a) $dx_i/dt = g_i(\partial\pi_i/\partial x_i)$, $g_i(0)=0$, $g^i_i>0$,

(8b) $dx_i/dt = h_i(y_i-x_i)$, $h_i(0)=0$, $h^i_i>0$,

where $y_i \equiv \psi^i(x_{-i})$ is the solution of

(9) $f^i(y_i,x_{-i})+y_i f^i_i(y_i,x_{-i})-C'_i(y_i) = 0$,

Assumption 6.

(10) $\pi^i_{ii}+\sum_{j\neq i}|\pi^i_{ij}| < 0$, $i=1,2,\cdots,n$.

Under this assumption and $g^i_i(0)>0$, it is known that Jg, the Jacobian matrix of the adjustment system (8a), becomes a matrix with a negative dominant diagonal. Hence the global stability of the Cournot oligopoly equilibrium.

Consider next (8b). The Jacobian matrix of (8b) has a negative dominant diagonal if and only if

$$1 > \sum_{j\neq i} |\phi^i_j| = \sum_{j\neq i} -|\pi^i_{ij}|/\pi^i_{ii}, \quad i=1,2,\cdots,n,$$

Hence the Cournot oligopoly equilibrium is globally stable under Assumption 6 for (8b). This establishes the stability-wise equivalence of the two adjustment systems (8a) and (8b). Note that because of the boundary condition, outputs remain nonnegative throughout adjustment periods.

Rewrite (10) for $i=1,2,\cdots,n$ as

$$(11.1)\quad 1 + \sum_{j\neq i} |f^i_j + x_i f^i_{ij}|/(2f^i_i + x_i f^i_{ii} - C''_i) > 0.$$

It is clear that (11.2) is satisfied if

$$(11.2)\quad |f^i_j + x_i f^i_{ii}|/(2f^i_i + x_i f^i_{ii} - C''_i) > -1/(n-1).$$

Let $MR^i \equiv \partial(x_i f^i)/\partial x_i$ be the marginal revenue (MR) of the i-th firm with respect to change in its output, $M^i_j \equiv \partial MR^i/\partial x_j$. Rewrite (11.1) as

$$(11.3)\quad MR^i_i + \sum_{j\neq i} |MR^i_j| < C''_i, \quad i=1,2,\cdots,n.$$

If the demand functions are all linear, and if, all goods are substitutes and $C''_i=0$, (11.1) reads

$$(11.4)\quad f^i_i < \sum_{j\neq i} f^i_j, \quad i=1,2,\cdots,n.$$

IV. ALTERNATIVE ADJUSTMENT SYSTEM

Let all assumptions in Section 3 are valid other than Assumption 5 and 6. Outputs might become negative during adjustment periods in the absence of Assumption 5 if they change according to (8a) or (8b). The following adjustment system is introduced to avoid this possibility.

$$(12)\quad dx_i/dt = \max(x_i + sk_i g^i(x), 0) - x_i, \quad i=1,2,\cdots,n,$$

where s and k_i's are positive constants and $g^i(x) \equiv \partial\pi^i/\partial x_i$ for all i.

Assumption (6) is now replaced with

Assumption 7. $g(x) = (g^1(x),\cdots,g^n(x))^T$, $x \in K \equiv \times[0,\bar{x}_i]$ is strictly monotone.

This assumption is equivalent to

$$(13)\quad (x^1-x^2)^T(g(x^1)-g(x^2)) < 0, \quad x^1\neq x^2, x^1, x^2 \in K.$$

The proof of the stability consists of two steps.

Step 1: Let x^* be the unique interior equilibrium, and let a Lyapunov function be given by

$$(14)\quad V(x) \equiv (x-x^*)^T K(x-x^*), \quad K=\mathrm{diag}(k_1^{-1},\cdots,k_n^{-1}).$$

Let $U \in K$ be an open neighborhood of x^* with sufficiently small radius. From $x^* + sK^{-1}g(x^*) > 0$ and continuity of g,

$$x + sK^{-1}g(x) > 0, \quad x \in U,$$

$$dx/dt = sK^{-1}g(x), \quad x \in U.$$

$$dV/dt = s(x-x^*)^T(g(x)-g(x^*)) < 0, \quad x\neq x^*, \ x \in U.$$

Step 2: There exists a positive number L such that $\sum_i k_i g^i(x)^2 \leq H$. Let

$$\beta \equiv \sup_{x\in K\setminus U} (x-x^*)^T(g(x)-g(x^*)) < 0.$$

From (12),

$$\dot{x}_i \equiv dx_i/dt \geq x_i + sk_i g^i(x) - x_i = sk_i g(x),$$

$$x^{*T}K\dot{x} \geq s x^{*T} g(x).$$

$$x_i + \dot{x}_i = \max(x_i + sk_i g^i(x), 0) \geq 0,$$

$$x_i^2 + \dot{x}_i^2 + 2x_i\dot{x}_i \leq x_i^2 + s^2 k_i^2 g^i(x)^2 + 2sk_i g^i(x) x_i,$$

$$\dot{x}^T K \dot{x} \leq s^2 \sum_i k_i g^i(x)^2/2 + s x^T g(x).$$

Hence taking into account $g(x^*)=0$, we evaluate dV/dt for $x \in K\setminus U$ as follows:

$$dV/dt = \dot{x}^T K \dot{x} - x^{*T} K \dot{x}$$

$$\leq s^2 H/2 + s\beta$$

$$= (H + 2\beta/s)s^2/2.$$

Since $\beta<0$, we conclude from this that $dV/dt<0$ if s is sufficiently small. Steps 1 and 2 together prove that $dV/dt<0$ for $x^*\neq x \in K$. Hence the Cournot equilibrium is globally stable for the adjustment system (12) if $s<-2\beta/H$.

REFERENCES

Al-Nowaihi, A. and P. L. Levine (1985), "The Stability of the Cournot Oligopoly Model: A Reassessment," Journal of Economic Theory, 35, 307-321.

Furth, D. (1986), "Stability and Instability in Oligopoly," Journal of Economic Theory, 40, 197-228.

Gandolfo, G. (1986), Economic Dynamics, rev. ed. North-Holland, Amsterdam.

Nikaido, H. and H. Uzawa, "Stability and Non-negativity of a Walrasian Tâtonnement Process," International Economic Review, 1, 50-59.

Okuguchi, K. (1964), "The Stability of the Cournot Oligopoly Solutions: A Further Generalization," Review of Economic Studies, 31, 143-146.

Okuguchi, K. (1976), Expectations and Stability in Oligopoly Models, Springer-Verlag, Berlin/Heidelberg/New York.

Okuguchi, K. (1983), "The Cournot Oligopoly and Competitive Equilibria as Solutions to Non-Linear Complementarity Problems," Economics Letters, 12, 127-133.

Okuguchi, K. (1990a), "Stackelberg and Cournot Duopolies Revisited," mimeo. Tokyo Metropolitan University.

Okuguchi, K. and F. Szidarovszky (1990b), "On the Uniqueness of Equilibrium in Cournot Oligopoly with Product Differentiation," Paper presented at the 6th World Congress of the Econometric Society, Barcelona, 22-28, August, 1990.

Okuguchi, K. and F. Szidarovszky (1990c), The Theory of Oligopoly with Multi-Product Firms, Springer-Verlag, Berlin/Heidelberg/New York.

Seade, J. (1980), "The Stability of Cournot Revisited," Journal of Economic Theory, 23, 15-27.

Szidarovszky, F. and S. Yakowitz (1982), "Contributions to Cournot Oligopoly Theory," Journal of Economic Theory, 25, 51-70.

THE IMPACT OF RISK AVERSION ON INFORMATION TRANSMISSION BETWEEN FIRMS

Yasuhiro Sakai

Institute of Social Sciences
University of Tsukuba
Tsukuba, Ibaraki 305, Japan

ABSTRACT This paper investigates whether and to what extent the presence of risk aversion affects the welfare implications of information transmission in one of the most fundamental oligopoly models ——— a Cournot duopoly model with common demand uncertainty. It aims to make a bridge between the literature dealing with information sharing in oligopoly and the one discussing the firm under uncertainty. We can show that the average output level of each firm is quite sensitive to the type and amount of information, and that the presence of risk aversion has an effect of decreasing the welfare of firms, whence information transmission may sometimes be harmful rather than beneficial to risk averse firms. These results have some policy implications.

1. A Duopoly Model with Risk Aversion

We assume that the market demand functions of outputs are given by

$$p_1 = \tilde{\alpha} - \beta(x_1 + \theta x_2),$$
$$p_2 = \tilde{\alpha} - \beta(x_2 + \theta x_1).$$

x_1 and x_2 are substitutes, independent, or complements according as θ is positive, zero, or negative. $\tilde{\alpha}$ is a random variable whose distribution is normal with mean μ and variance σ^2.

Letting π_i denote the profit of firm i, we obtain

$$\pi_i = (\tilde{\alpha} - c_i - x_i - \theta x_j)x_i \quad (i,j = 1,2;\ i \neq j).$$

Each firm is assumed to maximize the expected utility of its profits. We suppose that firm i the following von Neumann-Morgenstern utility :

$$U_i(\pi_i) = a - b\exp(-R_i\pi_i) \quad (b > 0,\ R_i > 0;\ i = 1,2).$$

We focus on the following three types of information structures: (i) no information, written as η^O, where no firm has information about the demand parameter $\tilde{\alpha}$; (ii) nonsymmetric information, denoted by η^N, where firm 1 is informed of $\tilde{\alpha}$ but firm 2 is not so informed; and (iii) shared information, shown by η^S, where both firms can know $\tilde{\alpha}$.

We say that a pair (x^O_1, x^O_2) of output strategies is an equilibrium pair under η^O if for each $\tilde{\alpha}$,

$$x^O_i = \underset{x_i}{Arg\ Max}\ EU_i[\pi_i(\tilde{\alpha}, x_i, x^O_j)] \quad (i,j = 1,2;\ i \neq j).$$

an output pair $(x^N_1(\tilde{\alpha}), x^N_2)$ is said an equilibrium pair under η^N if for each $\tilde{\alpha}$,

$$x^N_1(\tilde{\alpha}) = \underset{x_1}{Arg\ Max}\ U_1[\pi_1(\tilde{\alpha}, x_1, x^N_2)]$$
$$x^N_2 = \underset{x_2}{Arg\ Max}\ EU_2[\pi_2(\tilde{\alpha}, x^N_1(\tilde{\alpha}), x_2)].$$

Similarly an output pair $(x^S_1(\tilde{\alpha}), x^S_2(\tilde{\alpha}))$ is called an equilibrium pair under η^S if for each $\tilde{\alpha}$,

$$x^S_i(\tilde{\alpha}) = \underset{x_i}{Arg\ Max}\ U_i[\pi_i(\alpha, x_i, x^S_j(\tilde{\alpha}))] \quad (i,j = 1,2;\ i \neq j).$$

2. Comparison of Expected Outputs

THEOREM 1

(i) If $\theta > 0$, then $Ex^N_1 > Ex^S_1 > x^O_1$ and $Ex^S_2 > x^N_2 > x^O_2$;
(ii) If $\theta = 0$, then $Ex^S_1 = Ex^N_1 > x^O_1$ and $Ex^S_2 > x^N_2 = x^O_2$;
(iii) If $\theta < 0$, then $Ex^S_1 > Ex^N_1 > x^O_1$ and $Ex^S_2 > x^O_2 > x^N_2$.

In Figure 1 points Q^O, Q^N, and Q^S represent Cournot-Nash equilibrium under η^O, η^N, and η^S, respectively.

To sum up, the production activities of risk averse firms are seriously affected by the type and amount of information acquisition and by the degree of substitutability between goods. This is in marked contrast with the risk neutral case in which the average outputs remain unscathed by any change of information regardless of the degree of technical substitution.

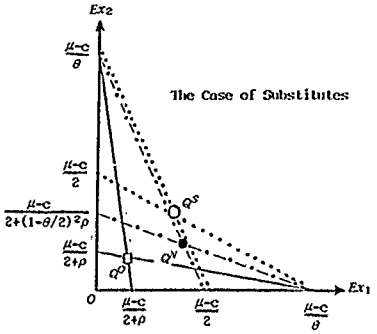


Figure 1. Comparison of Expected Outputs

3. Comparison of Expected Utilities

THEOREM 2 (η^O versus η^S)

$$EU^S_1 \gtreqless EU^O_1 \iff H(R,\sigma^2,\theta)\log(1+2\rho/(2+\theta)^2) \gtreqless (\mu-c)^2,$$

where $H(R,\sigma^2,\theta) = [2\rho+(2+\theta)^2][\rho+2+\theta]^2/(\theta R\sigma)^2$.

As is seen in Figure 2, the value of information is negative when $(\mu-c)$ is sufficiently large and/or when goods are either fairly strong substitutes or fairly strong complements.

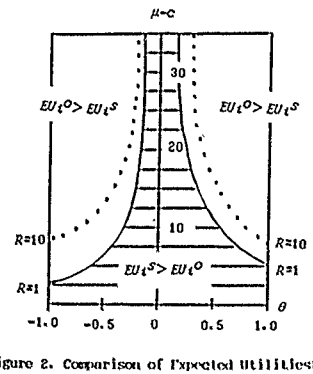


Figure 2. Comparison of Expected Utilities: No Information vs. Shared Information

THEOREM 3 (η^O versus η^N)

(i) $EU^N_1 \gtreqless EU^O_1 \iff J(R,\sigma^2,\theta)\log(1+\rho/2) \gtreqless (\mu-c)^2$,

where $J(R,\sigma^2,\theta) =$

$$\frac{(2+\rho)[(2-\theta)\rho+2(2+\theta)]^2(\rho+2+\theta)^2}{(\theta\rho)^2\{(\rho+2)[(2-\theta)\rho+2(2+\theta)]+(\rho+2+\theta)[(2-\theta)\rho+4]\}}$$

(ii) $EU^N_2 \lesseqgtr EU^O_2$, where the equality holds iff $\theta = 0$.

As was noted above, the anomalous situation under which EU^N_1 is less than EU^O_1 occurs if $(\mu-c)$ is sufficiently large and the absolute value of θ is fairly large.

THEOREM 4 (η^S versus η^N)

(i) $EU^S_1 \gtreqless EU^N_1$

$$\iff K(R,\sigma^2,\theta)\log\left\{1-\frac{\theta(4+\theta)\rho}{(2+\theta)^2(\rho+2)}\right\} \lessgtr (\mu-c)^2,$$

where $K(R,\sigma^2,\theta)$

$$= (-1)(2+\rho)[(2+\theta)^2+2\rho][(2-\theta)\rho+2(2+\theta)]^2/(\theta R\sigma)^2[(2+\theta)(2-\theta)\rho+8(1+\theta)].$$

(ii) $EU^S_2 \gtreqless EU^N_2$

$$\iff L(R,\sigma^2,\theta)\log\left\{1+\frac{2\rho}{(2+\theta)^2}\right\} \lessgtr (\mu-c)^2,$$

where $L(R,\sigma^2,\theta) = [2\rho+(2+\theta)^2][(2-\theta)\rho+2(2+\theta)]^2/(\theta^2 R\sigma)^2$.

Here again, the values of $(\mu-c)$ and θ play a vital role in deciding the welfare impact.

4. Concluding Remarks

We believe that the welfare results obtained so far have some policy implications regarding the effectiveness and limitations of industrial policies employed by the government when firms in an industry display strong risk aversion.

There remain some other directions in which our study may be further pursued.

REFERENCES

1. Sakai, Y. (1985): "The Value of Information in a Simple Duopoly Model," *Journal of Economic Theory* 36: 36-54.

2. Sakai, Y. (1987): "Cournot and Bertrand Equilibria under Imperfect Information," *Journal of Economics* 48: 213-232.

3. Sakai, Y., and Yamato, T. (1989): "Oligopoly, Information and Welfare," *Journal of Economics* 49: 3-24.

4. Sakai, Y., and Yoshizumi, A. (1989): "Risk Aversion and Information Transmission in a Duopolistic Market," Discussion Paper, University of Tsukuba.

THE CORE OF THE INDUCTIVE LIMIT OF A DIRECTED SYSTEM OF ECONOMICS.

Dave Furth
Faculty of Law, Dept of Economics
University of Amsterdam
OZ Achterburgwal 217-219
1012 DL AMSTERDAM
THE NETHERLANDS

Competetive markets are characterized by a large number of economic agents, at both (supply and demand) sides of the market and such that no agent has enough power to manipulate prices. That is all agents are price takers.
In the sequal we will define an economy through a set of agents. Each agent has a type, characterized by a vector of initial endowments (from a fixed, finite dimensional commodity space) and a continuous, monotone and concave utility function.
Aumann (1966) considered the case where the set of agents was a continuum and proved that for this case the set of Walras allocations and the Core coincide. This is the "Core-equivalence" theorem.
Nowadays this theorem is often proved by a proposition by Anderson (1978).
Consider a finite economy (when we speak of a finite, countable or uncountable economy those adjectives refer to the set of agents). We know that the Walras allocations are in the Core. For each core allocation we can determine the distance from the nearest Walras allocation. The supreme of those distances, taken over all core allocations, is the distance between the Core and the Walras equilibria. Among others, Anderson shows, that this distance goes to zero when the number of agents goes to infinity. This does not imply that in the limit the core and the Walras equilibria coincide, because both sets could be empty.
The existence of Walras allocations is easilly proved in the case of replica economies à la Debreu and Scarf (1963). This is due to the "equal treatment" property in the Core. All replica of a finite economy with a finite number of types share the "same" Walras equilibria. So also the limit economy has the same Walras allocations, that in the limit coincide with the Core.
For economies that are not replica of an economy with a finite number of types, the "equal treatment" property does not hold and above result is more difficult to prove. The way to prove the Core equivalence theorem in this case is with competitive sequences of economies (see Hildenbrand (1982)). One of the conditions is that the distribution (a counting measure) of types of agents in the finite economies converges to a distribution (a measure) of agents in the limit economy, that has a continuum of agents. Most proves of the core equivalence theorem for large economies use measure theory on a continuum of agents. As this world is finite, in my view, this is, with the continuum hypothesis, one step too many.
Approximations of large economies should be dealing almost with countable infinite agents sets.
Therefore I consider directed systems of (finite) economies. That is there is a directed index set, that is a set with a partial ordening (reflexive and transitive) such that for all two indices from that set there is a third larger than both of them. When of two indices, one is larger than the other, we have an embedding of the economy with the smaller index into the one with the larger.
The inductive limit is defined in the usual way. The limit economy may have a countable infinite agents set. The following Theorem can be proved:

Theorem: The Core of the inductive limit of a directed system of finite economies is non-empty.

This generalizes the Debreu-Scarf result above, as replica economies form a directed system of economies, but also the case of (non-replica) economies with a finite number of types can be treated this way. The remakable fact is that in the proof of above Theorem nu use is made of measure theory. As each economy with a countable infinite set of agents is the inductive limit of a directed system of finite economies, also the following Corrolary has been proved:

Corrolary: The Core of a countable infinite economy is non-empty.

Amsterdam, februari 1991

References:

Anderson, B.M. (1987): An elementery Core Equivalence Theorem, Econometrica, 46, p. 1483-1487.
Aumann, R.J. (1966): Existence of competetive equilibria on markets with a continuum of trades, Econometrica, 34, p. 1-17.
Debreu, G. and Scarf, H. (1963): A limit on the Core of an economy, International Economic Review, 4, p. 235-246.
Hildenbrand, W. (1982): Core of an economy, in: Arrow, K.J. and Intriligator, M.D. (editors): Handbook of Mathematical Economics, North-Holland, Amsterdam.

TOWARDS AN INTELLIGENT SCHEME FOR SOLVING LARGE-SCALE DYNAMIC MODELS WITH ECONOMIC APPLICATION [1]

YOSHIO KINERA
Faculty of Economics
Chukyo University
Showa-ku, Nagoya 466
JAPAN

and

HITOSHI KONDO
Faculty of Economics
Nanzan University
Showa-ku, Nagoya 466
JAPAN

Abstract In this paper, we propose a computing system which recognises the structural characteristics of a given large-scale dynamic model and solves the model through a programme automatically generated on the basis of recognised structural characteristics. Throwing light on the causal relationship which is inherent in the model, this kind of system is also useful for revising the model so as to improve its dynamic behaviour particularly when the model is not only of large-scale but of high complexity.

I. INTRODUCTION

In spite of the recent rapid progress in computing ability, there remain various difficulties in solving a large-scale dynamic model. For instance, a mechanical application of standard approximation-method is apt to render the iterative process endless even if the process is to be convergent or to result in inaccurate and unliable numerical solutions. However, as was asserted in a series of works [1] through [4], the above difficulties are known to be overcome, at least partly, by utilising the qualitative information of the model such as the structural knowledge on the model and/or the causal relationship of variables contained in the model. Furthermore, to our knowledge, the works cited (for example CAUSOR) confine themselves to the analysis of interdependent causal structures. Therefore it may be of some value to develop a computing system which numerically solves the given dynamic model with the aid of the qualitative knowledge perceived by the system itself.

1) This work is supported by a grant of Japan Ministry of Education (grant No. 01300002).

A. Symbol

We list the symbol to be used.

y : An $n\times 1$ vector of endogenous variables.

x : An $n\times 1$ vector of predetermined variables. If necessary, we specify the first n_1 elements of x as the lagged endogenous variables.

$f_i(y,x)$: The i-th equation of the system to be considered ($i=1,\cdots,n$).

$A=(a_{ij})$: The Jacobian matrix of the model under consideration. More specifically,

$$a_{ij} = \begin{cases} \partial f_i/\partial y_j & j = 1,\cdots,n \\ \partial f_i/\partial x_{j-n} & j = n+1,\cdots,n+n \end{cases}$$

A_{PQ} : A submatrix of A consisting of a_{ij} such that $i\in P$ and $j\in Q$, where P (Q) is any subset of all row-indices (column-indices)

y_P : A subvector obtained from y by extracting y_i such that $i\in P$

$N=\{1,\cdots,n\}$: The set of all indices of equations. However, N is replaced by V if we wish to emphasise that it is the set of all endogenous variables.

Throughout the paper, we are concerned with the following simultaneous system of dynamic equations;

$$f(y,x) = 0 \qquad (1)$$

, which we henceforth call the canonical form.

B. Terminology

Definition 1. System (1) is said to be decomposable if there exists a nonempty proper subset S of N such that corresponding to S there can be found a proper subset V_S of V with the subsequent properties:
(i) $\#S = \#V_S$

(ii) associated with any j of V_S there exists an i of S such that $a_{ij} \neq 0$, and

(iii) $a_{ij} = 0$ for any i of S and any j of $M-V_S$.

the complementary set of V_S with respect to M, where $\#S$ ($\#V$) denotes the number of elements contained in the set S (V).

In words, Definition 1 means that $\#S$ equations whose indices are in the set S contain just the same number of endogenous variables. Hence, if A_{SV_S} proves to be nonsingular, the numerical values of these endogenous variables are determined as the functions of predetermined variables. This, in turn, enables us to solve the remaining endogenous variables from the remaining equations. Concerning Definition 1, it should be noted that A_{MV} is a reducible matrix if and only if system (1) is decomposable with $S = V_S$.

Definition 2. We call system (1) partially linear if there exist nonempty subsets T of M and V_T of V such that A_{TV_T} is a constant matrix.

The concept of partial linearity is also useful for our purpose, since once the partial linearity is found, we can eliminate endogenous variables whose number equals the rank of A_{TV_T}. Needless to say, it does not harm generality to assume that $\#T$ does not exceed $\#V_T$, for otherwise the set T can be contracted until $\#T \leq V_T$.

II. THE COMPUTING SYSTEM

The proposed computing system is summarised as follows:

1. Rearrangement of the equations read into the canonical form. Needless to say, expressions which do no meet the FORTRAN77 grammar, if any, are pointed out before proceeding further.

2. Checking the decomposability and/or the partial linearity

2.1. In the case of decomposability, it is further checked whether or not the subsystem corresponding to the set S can be rewritten as

$$y_{V_S} = g(y_{\bar{V}_S}, x) \qquad (2)$$

, to which the socalled Gauss-Seidel method is directly applicable. If the Gauss-Seidel method is unapplicable, we resort to other methods, such as Newton method and, if necessary, the Jacobian matrix is generated through nonnumerical differentiation. The generation of the Jacobian matrix is executed by a programme written in PROLOG. [2]

2.2. If the partial linearity is found, some endogenous variables are eliminated to obtain the reduced (or contracted) system.

3. The generation of the computing programme in accordance with the computation formula specified in stage 2 above. Briefly stating, the programme for numerical solutions is generated by replacing the name of each variable with the corresponding location (array name) in the memory and by utilising the stored programming knowledge for carrying out the specified computational procedure.

REFERENCES

[1] Gabley, M. & M.H. Gilli (1984). "Two Approaches in Reading Model Interdependencies", Chapter 2 of Analysing the Structure of Econometric Model, ed., Ancot, J.P., The Hague, Nijhoff.

[2] Gallo, G.M. & M.H. Gilli (1984). "How to Strip a Model to Its Essential Elements", read at Colloque International ISMEA - Universite d'Ottawa.

[3] Gilli, M.H. & E. Rossier (1981). "Understanding Complex System", Automatica, vol. 17, 647-652.

[4] Gilli, M.H. (1984). "CAUSOR A Program for the Analysis of Recursive and Interdependent Causal Structure", Cahiers du Department d'Econometrie, Universite de Geneve.

2) We have defined the decomposability and the partial linearity in terms of the Jacobian matrix simply for the descriptive clarity Evidently, the decomposability, as well as the partial linearity, can be checked without having the knowledge on the Jacobian matrix Therefore, it is at this stage that the Jacobian matrix is constructed through nonnumerical differentiation.

A SURVEY OF POLICY OPTIMISATION ALGORITHMS

Berç Rustem
Department of Computing
Imperial College of Science, Technology & Medicine
180 Queen's Gate, London SW7 2BZ, United Kingdom

ABSTRACT

Central to most policy optimisation considerations is a deterministic optimisation algorithm. We discuss three such algorithms for nonlinear models. The first is a Gauss-Newton algorithm, ordinarily for unconstrained problems. It requires simple derivative information, it is slow but quite effective for rough computations. The second is a Goldstein-Levitin-Polyak type algorithm. It easily admits linear constraints such as bounds on the controls. The derivative information for this algorithm can be efficiently obtained using a backward evaluation approach. The algorithm is fast and involves a quadratic subproblem with inequality bounds. The third is a sequential quadratic programming algorithm that admits general equality and inequality constraints. The derivative calculations are expensive but the algoritm is fast and solves the most general formulation.

I. INTRODUCTION

Consider the policy optimisation problem

$$\min \left\{ J(Y, U) \,\middle|\, F(Y, U) = 0 \right\} \tag{1}$$

where Y and U are respectively the endogenous, or output, variables and policy instruments, or controls, of the system. J is the policy objective function and F is the model of the economy which is affected by the random disturbance vector ϵ. In general, F is nonlinear with respect to Y, U. Problem (1) is essentially a static transcription of a dynamic optimisation problem in discrete time where

$$U = \begin{bmatrix} u_1 \\ \vdots \\ u_t \\ \vdots \\ u_T \end{bmatrix}; \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_t \\ \vdots \\ y_T \end{bmatrix}$$

with $u_t \in R^n$ and $y_t \in R^m$ denoting the control and endogenous variable
vectors at time period t. The optimisation covers the periods t = 1, ..., T. Thus, $Y \in R^{m \times T}$, $U \in R^{n \times T}$, $F \subset R^{n_z} \to R^{m \times T}$ and $J: J \subset R^{n_z} \to R^1$ ($n_z = T \times (m+n)$). The vector valued function F is essentially an econometric model which is a system of nonlinear difference equations represented in static form for time periods t =1, .., T. The size of the system is generally in the order of hundreds of equations per time period. Problem (1) is unconstrained. We shall discuss a simple Gauss-Newton (G-N) type algorithm for its solution.

The second policy optimisation formulation we shall consider admits linear inequality constraints on the control variables

$$\min \left\{ J(Y, U) \,\middle|\, F(Y, U) = 0; N^T U \leq b \right\}. \tag{2}$$

We shall consider a Goldstein-Levitin-Polyak (GLP) algorithm for solving this problem.

To introduce the third problem, we shall use the vector x to denote

$$x = \begin{bmatrix} Y \\ U \end{bmatrix} \tag{3}$$

Consider the general nonlinear equality and inequality constrained problem

$$\min \left\{ J(x) \,\middle|\, F(x) = 0; G(x) \leq 0 \right\}. \tag{4}$$

We shall consider a sequential quadratic programming (SQP) algorithm for this problem.

II. G-N ALGORITHM FOR UNCONSTRAINED PROBLEMS

Associated with the model F (Y, U) = 0, there is a model solution that computes the value of Y, for a given value of U. The computational mapping is denoted by

$$Y = \xi(U). \tag{5}$$

This mapping can be used to eliminate the output variables from the problem. Hence, we can express (1) as

$$\min \left\{ J(U) \right\} \tag{6, a}$$

where

$$J(U) = J(\xi(U), U). \tag{6, b}$$

For this particularly simple algorithm, we also assume that the objective function is a quadratic and is given by

$$\tfrac{1}{2} \langle Y - Y^d, Q_y (Y - Y^d) \rangle + \tfrac{1}{2} \langle U - U^d, Q_u (U - U^d) \rangle$$

where Q_y, Q_u are symmetric matrices; Q_y is positive semi-definite and Q_u is positive definite, Y^d, U^d are given desired values. When the model is used to eliminate Y from this function, we have

$$Q(U) = \tfrac{1}{2} \langle \xi(U) - Y^d, Q_y (\xi(U) - Y^d) \rangle + \tfrac{1}{2} \langle U - U^d, Q_u (U - U^d) \rangle$$

Starting from an initial point U_0, the algorithm generates the sequence $\{U_k\}$, k=0, 1, ... where

$$U_{k+1} = U_k + \tau_k d_k.$$

The direction d_k is defined by

$$H_k^0 d_k = -\nabla Q(U_k) \quad \text{with} \quad H_k^0 = N_k^T Q_y N_k + Q_u$$

$$N_k = \frac{\partial \xi}{\partial U}\bigg|_{U = U_k} = \frac{\partial Y}{\partial U}\bigg|_{U = U_k}.$$

The Jacobian N_k is initially approximated by the dynamic multipliers of the model and subsequently updated using Broyden's rank-one formula

$$N_{k+1} = N_k + \frac{\{\xi(U_{k+1}) - \xi(U_k) - \tau_k N_k d_k\} d_k^T}{\tau_k \langle d_k, d_k \rangle}$$

in order to preserve the block lower triangular structure of the Jacobian, reflecting the causality-time structure of the model, Schubert's (1970) modification to the above formula can be used. The stepsize $\tau_k \in [0, 1]$ is determined using an Armijo strategy to satisfy

$$Q(U_{k+1}) - Q(U_k) \leq 10^{-4} \tau_k \langle \nabla Q(U_k), d_k \rangle$$

This algorithm is a generalisation of Holbrook's approach of linearising the model and solving the quadratic optimisation subject to the linearised model (Holbrook, 1974; Chow, 1975). Further deatails and properties of the algorithm are discussed in Rustem (1981). Although the algorithm is relatively slow, it is easy to implement on macro-models and requires little computational effort and has been successfuly used in policy optimisation exercises.

III. GLP ALGORITHM FOR LINEAR INEQUALITY CONSTRAINTS

The GLP algorithm is based on the projection of the unconstrained descent direction onto the convex feasible set of linear inequality constraints. Using the mapping (5), problem (2) can be written as

$$\min \left\{ J(U) \,\middle|\, N^T U \leq b \right\}. \tag{7}$$

The algorithm generates a sequence of points $\{U_k\}$ such that

$$U_{k+1} = U_k + \tau_k (\hat{U} - U_k)$$

where U is the projection of the unconstrained quasi-Newton step on the convex constraint set. The unconstrained step is given by

$$\mathcal{U} = U_k - \pi_k \hat{H}_k^{-1} \nabla J(U_k); \quad \pi_k \in [\underline{\pi}, \bar{\pi}] \in [1, 2)$$

where $\hat{H}_k$ is a positive definite quasi-Newton approximation to the Hessian of J at U_k and the unconstrained stepsize π_k is any sequence with the above restrictions and that $\{\pi_k\} \to 1$ (e.g. $\pi_k = 1, \forall k$). The projection of $\mathcal{U}$ on the constraint set is given by

$$\hat{U} = \arg\min \left\{ \|U - \mathcal{U}\|^2_{\hat{H}_k} \mid N^T U \leq b \right\}$$

and the stepsize $\tau_k \in [0, 1]$ is determined using an Armijo strategy such that the following merit function is satisfied

$$J(U_{k+1}) - J(U_k) \leq \tau_k \rho < \nabla J(U_k), \hat{U} - U_k >, \quad \rho \in \left(0, \tfrac{1}{2}\right)$$

The advantages of this algorithm are that ∇J can be efficiently computed using a back substitution scheme. In addition, the stepsize τ_k converges to unity and the algorithm converges to the solution at a Q-superlinear rate (Rustem, 1984).

IV. SQP ALGORITHM FOR NONLINEAR CONSTRAINTS

Consider definition (3) and the augmented Lagrangian of (4):

$$\mathcal{L}(x, \lambda, \mu, c, \alpha) = \ell(x, c, \alpha) + F(x)^T \lambda + G(x)^T \mu \tag{8}$$

where $0 \leq c \in R_+$, $0 \leq \alpha \in R^j_+$ and λ, μ are the Kuhn-Tucker multipliers. We define $(\cdot)_+$ such that its j th element, $(\cdot)^j_+ = \max\{(\cdot)^j, 0\}$. The penalty function associated with (3) is defined by

$$\ell(x, c, \alpha) = J(x) + \tfrac{c}{2} \| F(x) \|_2^2 + \tfrac{1}{2c} \| (c\, G(x) + \alpha)_+ \|_2^2. \tag{6}$$

Neither the augmented Lagrangian (5) nor the penalty function (6) possess continuous Hessians. In order to overcome the difficulties arising from this, we shall, at appropriate junctures, consider the effect of replacing the Hessian of (6) by that of

$$\hat{\mathcal{L}}(x, \lambda, \mu, c, \alpha) = \hat{\ell}(x, c, \alpha) + F(x)^T \lambda + G(x)^T \mu \tag{7}$$

where

$$\hat{\ell}(x, c, \alpha) = J(x) + \tfrac{c}{2} \| F(x) \|_2^2 + \tfrac{1}{2c} \| (c\, G(x) + \alpha)_+ \|_2^2 \tag{8}$$

$\hat{\ell}$ and ℓ differ only in their last terms corresponding to the inequality constraints. In contrast to (5), the Hessian of $\hat{\mathcal{L}}$ exists, provided J, F and G are suitably differentiable.

The successive quadratic programming (SQP) quasi-Newton algorithm considered in this paper, involves the basic iterative procedure

$$x_{k+1} = x_k + \tau_k d_k \tag{9}$$

where d_k solves the *Quadratic Programming Subproblem* (QPS)

$$\min \left\{ q(d, c_k, \alpha_k(x_k)) \;\middle|\; \nabla F_k^T d + F_k = 0, \;\; \nabla G_k^T d + G_k \leq 0 \right\} \tag{10}$$

the multipliers of this QPS, corresponding to d_k, are denoted by λ_{k+1}, μ_{k+1}. It is assumed that the feasible set of (10) is nonempty and that the QPS does have a solution. The objective function is given by

$$q(d, c, \alpha(x_k)) = d^T \nabla \ell(x_k, c, \alpha(x_k)) + \tfrac{1}{2} d^T \hat{H}_k d \tag{11}$$

with $\nabla \ell = \nabla_x \ell$. $\hat{H}_k$ is a quasi-Newton approximation to the Hessian of $\hat{\mathcal{L}}$, with respect to x, at x_k, λ_k, μ_k, c_k, α_k. Following the original suggestion by Wilson (1963), several variants of the SQP algorithm have been proposed. These have proved to be effective in solving problem (3) (see, e.g. Fan, Sarkar and Lasdon, 1988; and the references in Rustem, 1990).

Let the j th element of the vectors G and α be respectively denoted by G^j and α^j. We are mainly concerned with two specific values of α where α_0 is set in Step 0 with $c_0 \in (0, \infty]$ given as input. These are

$$\alpha^j_{k+1}(x_k) = \begin{cases} -c_{k+1} G^j(x_k) & \text{if } G^j(x_k) \leq 0 \\ \alpha^j_k(x_k) & \text{if } G^j(x_k) > 0 \end{cases} \tag{12,b}$$

and

$$\underset{\text{Algorithm}}{\alpha^j_{k+1}(x_{k+1})} = \begin{cases} -c_{k+1} G^j(x_{k+1}) & \text{if } G^j(x_{k+1}) \leq 0 \\ \alpha^j_{k+1}(x_k) & \text{if } G^j(x_{k+1}) > 0 \end{cases} \tag{12,c}$$

Given x_0, $c_0 \in R_+$, $\delta \in (0, \infty)$, $\rho_1 \in (0, 1 - \frac{1}{2c_1})$, $c_1 \in (\frac{1}{2}, 1]$, $\gamma \in (0, 1)$, $\hat{H}_0$.

Step 0: Set $k = 0$, $\alpha^j_0 = \max\{-c_0 G^j(x_0), 0\}$, $j = 1, .., j$.

Step 1: Solve (10) to obtain d_k and λ_{k+1}, μ_{k+1}.

Step 2: If optimality is achieved, stop. Else go to step 3.

Step 3: If

$$\nabla J_k^T d_k + c_1 d_k^T \hat{H}_k d_k - (G_k)_+^T \alpha_k(x_k) - c_k \{ \| F_k \|_2^2 + \| (G_k)_+ \|_2^2 \} \leq 0 \tag{13}$$

then $c_{k+1} = c_k$. Else set

$$c_{k+1} = \max \left\{ \frac{\nabla J_k^T d_k + c_1 d_k^T \hat{H}_k d_k - (G_k)_+^T \alpha_k(x_k)}{\| F_k \|_2^2 + \| (G_k)_+ \|_2^2}, \;\; c_k + \delta \right\} \tag{14}$$

Calculate $\alpha_{k+1}(x_k)$ using (12,b).

Step 4: Determine τ_k with $x_{k+1} = x_k + \tau_k d_k$ satisfying the inequality

$$\ell(x_{k+1}, c_{k+1}, \alpha_{k+1}(x_{k+1})) - \ell(x_k, c_{k+1}, \alpha_{k+1}(x_k)) \leq \tau_k \rho_1 \nabla \ell(x_k, c_{k+1}, \alpha_{k+1}(x_k))^T d_k \tag{15}$$

$\alpha_{k+1}(x_{k+1})$ is computed using (12,c).

Step 5: Update $\hat{H}_k$ to compute $\hat{H}_{k+1}$.

Step 6. Set $k = k + 1$ and go to Step 1.

The above algorithm allows the most general formulation. The stepsize τ_k converges to unity, the penalty parameter c_k does not grow indefinitely and $\{x_k\}$ converges to the solution at a two-step Q-superlinear rate. However, its computational demands in terms of derivative evaluations are considerable for large scale problems.

ACKNOWLEDGEMENTS

The financial support of ESRC is gratefully acknowledged.

REFERENCES

1. Chow, G.C. (1975). *Analysis and Control of Dynamic Economic Systems*, John Wiley, New York.
2. Fan, Y, S. Sarkar, L. Lasdon (1988). "Experiments with Successive Quadratic Programming Algorithms", *JOTA*, 56, 359-382.
3. Holbrook, R.S. (1974). "A Practical Method for Controlling a Large Nonlinear Stochastic System", *Annals of Economic and Social Measurement*, 3, 796-811.
4. Schubert, L.K. (1970). "Modification of a Quasi-Newton Algorithm for Nonlinear Equations with a Sparse Jacobian", *Mathematics of Computation*, 24, 27-30.
5. Rustem, B. (1981). *Projection Methods in Constrained Optimisation and Applications to Optimal Policy Decisions*, Springer Verlag, Berlin.
6. Rustem, B. (1984). "A Class of Superlinearly Convergent Projection Algorithms with Relaxed Stepsizes", *Appl. Math Optim*, 12, 29-43.
7. Rustem, B. (1990). "Equality and Inequality Constrained Optimisation Algorithms with Convergent Stepsizes", *PROPE Discussion Paper 70*, Imperial College.
8. Wilson, R B. (1963). *A Simplicial Algorithm for Concave Programming*, Ph. D. Dissertation, Harvard University.

ON THE UNIQUENESS AND THE STABILITY OF THE STEADY STATES IN A COMPUTABLE GENERAL EQUILIBRIUM MODEL WITH OVERLAPPING GENERATIONS

Pierre-Yves LETOURNEL
Cuong LE VAN
Katheline SCHUBERT

Abstract.

In this paper we present a computable general equilibrium model with production. At any time, the household sector comprises two overlapping generations of adults. We obtain the following results :

i) in the case of no taxes and no governement expenditures, the model has two steady states associated respectively with a strictly positive and a null interest rate.

ii) the introduction of taxes yields two steady states with a strictly positive and a strictly negative interest rate.

iii) if the government expenditures are large, then one can have two steady states with strictly positive interest rates or no steady state.

v) the steady state with lowest interest rate is locally stable while the one with highest interest rate is unstable.

INTRODUCTION

The use of Computable General Equilibrium (CGE) models is now current practice. The early 1980's saw a "flurry" of these models dealing with three types of problems :

- Sectorial problems (energetic, agricultural)
- International trade
- Fiscal Policy

(see e.g Kemal Dervis et al. (1982), Shoven and Whalley (1984), Fullerton, Henderson and Shoven (1983)). Auerbach and Kotlikoff (1987) constructed a large-scale dynamic fiscal CGE model where consumers have a life-cycle behaviour and enterprises maximize their intertemporal profit. These agents are assumed to have perfect foresights on prices and wages. Their model includes also taxes and public expenditures. It is used to study dynamic fiscal policy. Mathematically two kinds of problems arise from this model :

i) the existence and the uniqueness of the steady states ;

ii) the stability of these steady states.

It is well-known that, in a two periods overlapping generations model without production, taxes and public expenditures, there exist two steady states ; with one of them is associated a positive interest rate, while with the other one is associated a zero interest rate (see, e.g Benassy and Blad (1990)).

The problem of the existence and the uniqueness of the steady states has not been studied by Auerbach and Kotlikoff. In order to study the stability of the steady states, Laitner (1990) points out that the model of Auerbach and Kotlikoff is similar to rational expectation (RE) models. Therefore, if around a steady state the number of stable eigenvalues is equal to the number of initial conditions, there exists, under some regularity conditions, an unique transition path which asymtotically converges to this steady state. But, implicitly, on this path, the consumers born before the date of shocks (policy changes) are constrained.

The main purpose of this paper is, using a very simple model with overlapping generations, production, taxes and public expenditures, on the one hand to study the existence and the uniqueness of the steady states and, on the other hand, to introduce a definition of transition paths where the consumer born before the date of shocks could be satisfied. The RE transition path will be a particular case of these paths.

Concerning the steady states we have the following result: if the government expenditures are less than some amount of taxes then there exist always two steady states associated respectively with a positive and a negative interest rate. If these expenditures are more than this amount, first, one has two steady states with positive interest rates and beyond some value of the expenditures, one has no steady state.

Concerning the transition path, we assume that the consumer born before the date of the policy changes has a desired consumption for this date. The equilibrium transition path will be defined as the one where every agent is satisfied, in particular, the consumer born before the date of shocks. The RE path can be viewed as a particular path associated with some desired consumption of this consumer. Then we have the following result: for most of the desired consumption functions, the steady state with negative interest rate (or the smallest positive interest rate) is locally stable while the one with positive interest rate (or greatest positive interest rate) the steady state of the RE path is locally unstable.

1. The model

In this model there are two goods, labor and a physical good which can be consumed or used as capital good. The consumer lives two periods. In every period a new trader is born ; therefore at each date there are exactly two consumers, a young and an old one. The economy begins at date 1. There is an entreprise which maximises its intertemporal profit from date 1 to infinity. We assume that labour is exogenously fixed. The agents are assumed to have perfect foresight.

1.1 - The consumer

Let $c_{1,t}$, $c_{2,t}$ denote the consumptions at periods t and t+1 of the agent born at date t. He (she) maximises his (her) intertemporal utility, $u(c_{1,t}) + \delta' u(c_{2,t})$, under the constraint

$$p_t c_{1,t} + p_{t+1} c_{2,t} \leq W_t(1-\bar{l})(1-\tau_t) + W_{t+1}(1-\bar{l})(1-\tau_{t+1})$$

Where p_t, p_{t+1} and W_t, W_{t+1} are prices and nominal wages at date t and t+1, τ_t, τ_{t+1} are tax rates at the same dates ; $\bar{l}$ is exogenous leisure.

1.2 - Production

There is one producer who maximises his (her) intertemporal profit

$$\max \sum_{t=1}^{\infty} (p_t y_t - p_t I_t - W_t L_t)$$

under the constraints

$$I_t = K_{t+1} - (1-\delta) K_t$$

$$Y_t = F(K_t, L_t)$$

$$K_1 \geq 0 \text{ given}$$

Where y_t, K_t, I_t L_t denote production, capital stock, investment and labor.

1.3 - Markets clearing

Let (G_t) denote the sequence of public expenditures. Then one has, for $t \geq 1$:

$$(5) \quad c_{1,t} + c_{2,t-1} + K_{t+1} - (1-\delta) K_t + G_t = F(K_t, L_t)$$

$$(6) \quad L_t = 2(1-\bar{l})$$

1.4 - Equilibrium

At date 1, relation (5) is fulfilled if one knows how to determine $c_{2,0}$, the consumption at that date of the agent born at date 0.

First, this agent has to face his (her) budget constraint

$$c_{2,0} = \omega_1(1-\bar{l})\mu_1 + \frac{E_0}{\beta_0}$$

where E_0 is his (her) savings transferred from date 0 to date 1.

Secondly one can assume that this agent has a desired consumption at date 1, $c^*_{2,0}$, which depends on ω_1, μ_1, β_0 and his (her) consumption at date 0, $c_{1,0}$. One can write :

$$c^*_{2,0} = \Psi(c_{1,0}, \omega_1, \mu_1, \beta_0)$$

Therefore, our economy has initial data $(K_1, c_{1,0}, E_0)$.

1.5 - Steady state

Assume now that the sequences (G_t), (τ_t) are stationary :

$$G_t = G, \tau_t = \tau, \qquad \forall t \geq 1$$

A steady state associated with there sequences is a triple $(\bar{K}, \bar{c}_1, \bar{E})$ such that the equilibrium from it verifies

$$\forall t \geq 1 \quad , K_t = \bar{K} ; c_{1,t} = \bar{c}_1 ; E_t = \bar{E}$$

where $E_t = \omega_t(1-\bar{l})\mu - c_{1,t}$

with $\mu = 1 - \tau$

It can be easily checked that the associated function Ψ has the following form :

$$\Psi(c_{1,0}, \omega_1, \mu_1, \beta_0) = c_{1,0}\, \phi\left(\frac{\delta'}{\beta_0}\right)$$

1.6 - Transitory path

Given a steady state $(\bar{K}_1, \bar{c}_1, \bar{E})$, a transitory path from this steady state and associated with sequence (G_t) and (τ_t) is an equilibrium with initial data $(\bar{K}_1, \bar{c}_1, \bar{E})$.

P.Y.LETOURNEL ; Direction de la Prévision - Paris - France
C.LE VAN ; CNRS-CEPREMAP, 140, rue du Chevaleret - 75013 - Paris - France
K. SCHUBERT ; Université de Tours - France

SOLVING DYNAMIC MODELS UNDER CONSISTENT EXPECTATIONS WITH MARKET EQUILIBRIUM CONDITIONS

ELIAS DINENIS
Centre for Economic Forecasting
London Business School
London NW1 4SA
England

SEAN HOLLY
School of Management and
Economic Studies
PO Box 595
Crooksmoor Building
Conduit Road
Sheffield S10 1FL
England

Abstract We discuss some the algorithmic approaches to the solution of econometric models which contain both forward looking expectations and market clearing equilibrium conditions. We find in favour of first order iteration methods as being the most robust method of model solution, and report some simulations on the LBS model.

Introduction

The application of numerical methods and the use of optimal control techniques in economics has been transformed over the last decade by the impact of rational expectations (see Holly and Hughes Hallett, 1989). There is now a variety of methods which can be used to solve econometric models in which expectations are forward-looking or rational. There is the multiple shooting technique of Lipton et al (1982), the extended path method of Anderson (1979), and Fair and Taylor (1983), and refinements of this approach introduced by Hall (1985) and Fisher et al (1986). Finally there is the penalty function method of Holly and Zarrop (1983).

Recently there has also been a increase in the use of equilibrium, market clearing frameworks, especially for modelling financial markets. In this paper we consider how methods for solving nonlinear rational expectations models can be extended to allow asset prices to clear spot markets – so demand is equal to supply – as well as allowing expectations to be forward looking.

Solution Methods

A deterministic nonlinear model with expectations of next periods's endogenous variables can be written.

$$f_i(y_t,...,y_{t-s},y^e_{t+1},x_t,...,x_{t-r}) = 0 \qquad (1)$$
$$i=1,...,m; \qquad t=1,...,T$$

Where y_t is a m–vector of endogenous variables and x_t is a n–vector of exogenous variables or policy instruments. s and r denote maximum lags. For compactness we can rewrite (1) as.

$$F(Y,Y^e,X) = 0 \qquad (2)$$

Where $Y = (y^1_1,y^1_2,...,y^1_t,...,y^1_T)$, $Y^e = (y^e_2,y^e_3,..y^e_{t+1},..y^e_{T+1})$, and $X = (x^1_1,x^1_2,...,x^1_t,..x^1_T)$. In order for the problem to be well defined we need a terminal condition for y^e_{T+1}, which we shall discuss later. A penalty function method simply requires the minimisation of the quadratic loss function.

$$J(Y,Y^e,X) = 1/2[(Z'QZ)+(Y^{e'}NY^e)] \qquad (3)$$

where $Z = (z^1_1,z^1_2,..z^1_t,..z^1_T)$, and $z_t = y^e_t - y_{t+1}$, for t=1,T–1, and $z_T = y^e_T - \varphi$, where φ is the terminal condition Q is a semi–definite positive weighting matrix, and N is a positive definite weighting matrix. Clearly the unconstrained minimum of (3) implies that the solution lies on the saddlepath and that z_t=0, for t=1,T.

When the model also contains equilibrium conditions that in asset markets supplies are always equal to demand and the asset price ensures that this holds, then it is possible to extend the penalty function method by defining $q_t = d_t - s_t$, where d is demand and s supply. Then the addition of the quadratic terms Q'KQ and P'RP will enable the set of prices, $P = (p^1_1,p^1_2,..,p^1_t,..p^1_T)$ to ensure Q=0, where K is a semi–definite positive weighting matrix and R is a positive definite weighting matrix.

In practice as with most applications of penalty function methods the dimension of the problem can cause difficulties. If a large number of expectational variables and asset prices appear in the model then the size of the problem can become unmanageable. This suggests that alternative methods which do not require the direct minimisation of (3) and the calculation of the first and second derivatives may be preferable.

One such method is suggested by the used of generalised first order iteration methods With these methods the process of solution involves an inner and an outer loop. Also within the inner loop there is a further loop of the form:

$$p^{k+1}_t = p^k_t (d_t/s_t)^{1/2} \qquad (5)$$

where k is the iteration counter, and p_t is a vector of asset prices which clear the market for financial assets. The use of a square root proved to speed up convergence. Within the inner loop there is an extrapolation parameter (λ_0) and also an extrapolation parameter for the outer loop (λ_1). In the Fair–Taylor method the inner loop iterates to convergence before using the outer loop. In the method of Hall(1985) only one iteration is used in the inner loop pass. However, we have found that confining the inner loop to a single iteration drove the model outside the feasible solution region, so the number of inner loop iterations was restricted to 4. This also had the effect of restricting the iteration count of (5).

Some Empirical Illustrations.

We used the LBS econometric model (see Dinenis et al (1989) for a description) in order to carry out some comparisons of different methods of solving models with equilibrium conditions and forward looking expectations. In the LBS model there are three market clearing prices for the equity market, the gilts market and the foreign exchange market. Lying behind each of these markets is a dis–aggregated set of asset demand and supply relationships for different sectors of the

domestic economy as well as overseas. Corresponding to each of these prices is also an expected price. Strictly speaking for the foreign exchange market it is the exchange rate which is used. Also the degree of substituteability between domestic and foreign assets governs the extent to which an uncovered interest parity condition of the form:

$$e_t = e^e_{t,t+1} + r_t - rsw_t \qquad (6)$$

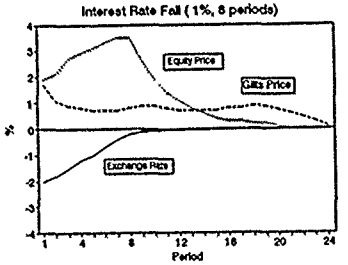
actually holds, where e is the exchange rate, r the domestic, and rsw the foreign interest rate, and $e^e_{t,t+1}$ the expectation of the exchange next period conditioned on information at time t. We found that the degree of substitution between domestic and foreign assets was implausibly low, and we therefore imposed the UIP condition given by (6). Thus the number of market clearing prices was reduced to two.

In Chart 1 we show the effects on all three asset prices of a temporary cut in short term interest rates of 1 percentage point for eight quarters. For the simulation the inner loop extrapolation parameter was set to 0.7, and the outer loop extrapolation parameter was 0.8 for the exchange rate, 0.8 for the price of gilts and 0.7 for the price of equity. The particular form of (6) means that a temporary fall in interest rates generates an immediate downward jump in the exchange rate in order for the exchange rate to then rise just enough to produce a capital gain equivalent to the fall in interest rates.

However, because an explicit arbitrage relationship does not determine the behaviour of the equity and gilts prices, we rely on the structural asset supply and demand equations to provide the degree of substituteability. Interestingly, both equities and gilts jump immediately in response to the unanticipated fall in interest rates by about 2%. However, thereafter they diverge. The equity price continues to rise peaking in the eigth quarter, and then declining back to zero as the interest rate rises back to its original level. However, the gilts price comes back only slowly, implying considerable imperfect substitutability.

CHART 1

Interest Rate Fall (1%, 8 periods)



References

Anderson P.A., "Rational expectations forecasts from non-rational models", Journal of Monetary Economics, 1979

Dinenis E., Holly S., Levine P., and Smith P., "The London Business School econometric model: some recent developments". Economic Modelling, July 1989.

Fair, R.C. and Taylor, J., "Solution and maximum likelihood estimation of dynamic nonlinear rational expectations models", Econometrica, 1983.

Fisher P., Holly S., and Hughes Hallett, A., "Efficient solution techniques for dynamic rational expectations models", Journal of Economic Dynamics and Control, 1986.

Hall, S., "On the solution of large economic models with consistent expectations", Bulletin of Economic Research, 1985.

Holly S., and Zarrop M., "On optimality and time consistency when expectations are rational", European Economic Review, 1983.

Holly S and Hughes Hallett,A., Optimal Control Expectations and Uncertainty, Cambridge University Press, 1989.

Lipton D., Poterba, J., Sachs J. and Summers L., "Multiple shooting in rational expectations models", Econometrica, 1982.

ON DETERMINISTIC AND STOCHASTIC SIMULATIONS OF NONLINEAR RATIONAL EXPECTATIONS MODELS

BOUCEKKINE R. CREST-CEPREMAP
142 rue du Chevaleret 75013 - Paris France

ABTRACT: This contribution aims at presenting some methods for solving nonlinear rational expectations models. These methods are applied on a model of the French economy, built up by Laffargue, Malgrange and Pujol (1990). At first, a NEW METHOD, introduced by Laffargue(1990), is used for deterministic simulation of the whole model. This method is a relaxation one based on the Newton-Raphson algorithm, contrary to the usual frameworks. Then, another method is conducted, consisting of a backward mapping stochastic simulation, using a normality assumption. Finally, it is shown how the previous assumption can be relaxed: especially, the method of approximation introduced by Tauchen(1986) is discussed.

I. THE MODEL:

The model considered, called PLM, is due to Laffargue, Malgrange and Pujol (1990). Wage's nominal rigidity and monopolistic competition in the good's market (based on an arbitrage between the aggregated domestic good and the imported one) are associated to generate Keynesian multipliers; the expectations of the agents are assumed to be rational. The production block uses a C. E. S production function and includes an adjustment cost for the capital. A constrained intertemporal optimization of the aggregated domestic firm, achieved by the Bellman method, provides two Euler equations. The consumption block is solved by the same method and provides one Euler equation. The model is "closed" by introducing the national government's expenditures and taxs, and by specifying some simple rules of its foreign economic policy.

The model is considered in a reduced, calibrated form; it can be decomposed in two inequal parts: an interdependent block of fourteen equations including the three Euler equations described above, and an epilogue of five equations. The first block can be seen as a dynamic system determining at each date t, the values of fourteen endogenous variables, using the lagged values of five predetermined ones and the one-period ahead values of five anticipated ones, thus four endogenous variables are static. Furthermore, this block contains six exogenous variables and is completed by giving initial values for the predetermined variables and terminal values for the anticipated ones.

II. LAFFARGUE'S METHOD:

Consider a model with a general form:

$f(y_1(t-1), y(t), y_2(t+1), L_t) = u_t$, for t=1 to T, where $y_1(t)$ is the $(n_1 x 1)$ vector of the model's predetermined variables, $y_2(t)$ is the $(n_2 x 1)$ of the anticipated ones, $y(t)$ is obtained by stacking $y_2(t)$, $y_3(t)$ and $y_1(t)$ in this order with $y_3(t)$ the $(n_3 x 1)$ vector of the static variables, u_t and L_t are respectively a white noise and the exogenous variables vectors. Furthermore, the vector of the initial conditions, $y_1(0)$, is given and we have a terminal condition of the form

$C(y_3(T), y_1(T), y_2(T+1))=c$, c is a constant vector.. Our purpose is to solve the constrained system described above. Given that f and C are respectively (nx1) and $(n_2 x 1)$ vectors of functions, with $n=n_1+n_2+n_3$, the system to solve has $nT+n_1+n_2$ equations and the same number of unknown variables. Put the following variable transformation. $x_1(t)= y_1(t)$, $x_2(t)=y_2(t+1)$, $x_3(t)=y_3(t)$. Denote by x(t) the vector obtained by stacking $x_3(t)$, $x_1(t)$ and $x_2(t)$ in this order, and by x'(t) the vector obtained by stacking $x_1(t)$ and $x_2(t)$ in this order. Thus the system can be written:

$f(x'(t-1), x(t), L_t)=0$, $t = 1$ to T, $S_0 x'(0)=x_1(0)$, $C(x(T)) = c$. S_0 is a well known matrix. The principle of Laffargue's method is the following, beginning with an initial solution path, the previous system is successively linearized using the Newton-Raphson algorithm. Denote by $x^{j-1}(t)$, t=0 to T, the solution path obtained at the step (j-1); the solution path at the step (j) is determined by:

$$x^j(t) = x^{j-1}(t)+\Delta x^j(t) \quad , \text{for } t = 0 \text{ to } T,$$

with $\Delta x^j = (\Delta x^j(0),.....,\Delta x^j(t))'$ solution of the linear system: $S^{j-1}\,\Delta x^j = s^{j-1}$; where S^{j-1} is the Jacobian matrix evaluated at the step (j-1) and s^{j-1} is a vector of constants. Let $K=nT+n_1+n_2$ For each j, S^j is a square (KxK) matrix ; Δx^j and s^j are (Kx1) vectors.

To solve the linear system, the matrix S^{j-1} is seen as a concatenation of (T+2) blocks; Gaussian elimination is conducted inside each block and when associated with other simple matricial operations, the whole matrix is transformed into a triangular form with each diagonal term equal to 1. Thus, Δx^j can be computed and a new solution path is calculated. The algorithm is like this reproduced until no improvement by linearization is possible.

One can see that the main problem is the invertibility of the matrices S^j; equivalently, the problem is to guarantee the unicity of the solution path:to ensure that, it is possible to transform the model,as it's explained in Laffargue (1990). In fact, it is more tractable to manipulate the model to ensure that the anticipated variables, in their advanced form, only appear in the same linear combinations, which is a sufficient condition for the unicity of the solution path. The interdependent block of PLM is solved using the previous result; the Blanchard-Kahn property is then checked.However, the configuration of the calculated eigenvalues impose to take high values for T -400.

III. BACKWARD-MAPPING METHODS:

The backward-mapping approach was suggested by Sims (1985). While the standard engineering method takes arbitrary stochastic processes for the exogenous variables to compute the corresponding endogenous processes,this approach suggests just the contrary. A meaningfull application of this method was done by Lee(1989),on a small standard stochastic equilibrium. Our view,here,is to see some numerical consequences of applying this method to more complicated and bigger models. At first,the model is simulated, keeping the normality assumption used by Lee,and the processes of the available solutions are analyzed.

A.-The method. The number of the endogenous variables' processes should be equal to the number of the exogenous variables of the considered block.thus,we can take six "entries". The processes of the available solutions are,then, explicitely computed without any approximation, given that only the endogenous variables are appearing in an anticipated form in the Euler equations of the PLM model. Precisely,we choose the "entries" so as the Euler equations can be easily solved. Let $(X_{i,t})$,with i=1,2, ...,6 and t the time index,be the chosen forms of "entries". Denote by x the logarithm of X. Assume,now,that the distribution of the stochastic vector $x_t=(x_{1,t},.....,x_{6,t})'$ is an autoregression of a given order p,natural integer, so as the following formula holds.

$$x_t=C(0)+C(1)\,x_{t-1}+C(2)\,x_{t-2}+. ..+C(p)x_{t-p}+u_t \qquad (1)$$

with.C(0) a (6*1) vector of constants and C(k) (k=1,2,..,p) (6*6) matrices of constants. Following Lee,we assume that the noises $(u_t)_t$ are not serially correlated and that u_t is normally distributed $N(m,\sigma)$, for each t, with $m=(m_1,m_2,. .,m_6)'$ a (6*1) vector and $\sigma=(\sigma_{i,j})$ the (6*6) variance-covariance matrix.

In order to generate dynamics as realistic as possible,the coefficients $C(k)$,$k=1$ to p,m and σ are estimated on a real data. Given these estimated coefficients,we can,at the beginning,compute the values of the processes $(X_{i,t})_{i,t}$,as follows:

$$\log(X_{i,t})=C_i(0)+\sum_{s=1}^{p}\sum_{j=1}^{6} C_{i,j}(s)\log(X_{j,t-s})+u_{i,t}$$

from (1),for each i,t.

Thus,the values of the processes $(X_{i,t})_{i,t}$ can be calculated:

$$X_{i,t}=\exp(C_i(0))\prod_{s=1}^{p}\prod_{j=1}^{6}(X_{j,t-s})^{C_{i,j}(s)}\exp(u_{i,t})$$

Given the p initial values $(x_1,x_2,.....,x_p)$ and generated random values from the law of u_t,the paths of each variable $X_{i,t}$ are easily computed. Moreover,the normality assumption, associated with a good choice of the auxiliary variables $X_{i,t}$, simplifies greatly the computations involved by the Euler equations; in effect, taking two reals a,b, we have the following:

$$E_t(X_{i,t+1})^a(X_{k,t+1})^b=\exp(\alpha)\prod_{s,j}(X_{j,t+1-s})^{\beta_{ik}(s,j)}\cdot\exp(\phi)$$

with for s=1 to p and j=1 to 6:

$\alpha=a\,C_i(0)+b\,C_k(0)$, $\quad\beta_{ik}(s,j)=a\,C_{i,j}(s)+b\,C_{k,j}(s)$

and $\phi=a\,m_i+b\,m_k+a^2\,\sigma_{ii}/2+b^2\,\sigma_{kk}/2+a\,b\,\sigma_{ik}$.

$E_t(.\ )$ is the expectation operator conditionally to the information set available at the date t. These formulas allow us to conduct the main computations involved by Euler equations.

B. Markov process approximation. If the exogenous variables appear in an anticipated form under the operator $E_t(.)$ or if we want to relax the normality assumption, we have,then, to develop an other framework. That is Tauchen's methodology which we have introduced in the context of this study. Unfortunately, this method, used as a backward-mapping one, is too numerically expensive if conducted on models including more than three or four exogenous variables,as it can be seen further. For a purpose of simplification, let 's consider a model with n exogenous variables ($n<5$); $(X_{i,t})_{i,t}$ are the auxiliary endogenous variables chosen for backward-mapping (then,$i=1$ to n). With the same notations as in C)-1,assume that the vector x_t satisfies the property (1) with $p=1$ but without the normality of u_t. We have, now,to approximate the AR(1) process by a Markov one with a finite number of states. Four steps must be achieved:

a) We transform the relation (1) to obtain an AR(1) process without intercept and with a diagonal variance-covariance matrix for the noises,which we note $\Omega=(\omega_i)$. $(X'_{i,t})$ are the transformed "entries".

b) The discrete-valued Markov process,denoted by $(X'_{i,t,d})$, approximating the AR(1) process, is then computed. Assume that each variable $X'_{i,t,d}$ has k_i values,given by:

$$X_{i,t,d}(k)=D_i(k-(k_i+1)/2),\ i=1\text{ to }n,\ k=1\text{ to }k_i$$

where: $D_i=2\,r\,\omega_i/(k_i-1)$; $r=3$ was found to work well. Consequently, the vector x'_t has N states, $N=k_1\,k_2...\,k_n$.

c) The transition matrix $\Pi=(\pi(s,s'))$,s and s'=1 to N, is then calculated, using the estimated coefficients of the transformed AR(1) process.

d) Finally,we reverse the transformation and return to the initial relation (1): the values of the Markov process are modified, but not the matrix Π.

By the Skorohod construction, the discrete-valued Markov process converges almost-surely to the AR(1) considered. Suppose, now,that we have an Euler equation of the general form: $z_t=E_t(g(x_{t+1},z_{t+1}))$, where z_t is a process to be computed and g a given function . Put the previous relation in an integral form:

$$z(x)=\int g(x'',z''(x))\,dF(x''/x)\quad\text{where } F(\,.\,)\text{ is the conditional}$$

cumulative probability distribution function of the process x_t. Use,now, the approximation $x_{t,d}$ of x_t: the integral can be discretised as follows:

$$z_{t,d}(s)=\sum_{s'=1}^{N}\pi(s,s')\,g(x_{t,d}(s'),z_{t,d}(s'))\ \text{ for s=1 to N}$$

The previous system has N equations and N unknown variables $z_{t,d}(s)$,s=1 to N; it's generally easily solved. The main difficulty of this method is that N is likely to be large,provoking high computational costs, that is why we have suggested that n must be small. To give an example of such a framework,we have considered the production block of PLM and we have transformed it so as it remains three exogenous variables (among five initially). Thus, n=3 and choosing $k_1=k_2=k_3=8$, we have N=512.The block solved includes two Euler equations, one in a quadratic form and the other is linear.

IV. CONCLUSION:

Regarding to the macroeconomic models curvature, Laffargue's method is likely to be numerically efficient: that's because we expect a small number of iterations of Newton-Raphson type. Using a formal derivator, the method is very tractable. In the other hand, if the model in its inital form does not ensure the unicity solution condition, it is generally possible to transform it, by simple operations, to get the required form.

Backward-mapping methods are, also, interesting for a purpose of stochastic simulation or estimation. Tauchen's method is found to be feasible only if the model to solve includes a small number of exogenous variables but even in this case, its computational cost is important. By the use of the normality assumption, the framework is numerically simplified; however, when simplifying the computations by a convenient choice of auxiliary variables, we may not obtain certain variables in their structural form.

References:

Laffargue, J-P (1990): "Résolution d'un modèle macroéconomique à anticipations rationnelles " *Annales d'Eco. et Stat.*, 17.

Laffargue, J-P, P. Malgrange and T. Pujol (1990): "Une maquette trimestrielle de l'économie française avec anticipations rationnelles et concurrence monopolistique " CEPREMAP Working Paper.

Lee, B. S (1989): "Solving, estimating and testing a nonlinear stochastic equilibrium model, with an example of the asset returns and inflation relationship " *J. E.D. C.*, 13.

Sims, C (1985), " Solving nonlinear stochastic equilibrium models backwards " *Discussion Paper 206* (University of Minnesota).

Tauchen, G (1986): "Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data" *Journal of Business and Economic Statistics*,4

A NEW ALGORITHM FOR OPTIMUM STOCHASTIC CONTROL OF NONLINEAR ECONOMIC MODELS*

JOSEF MATULKA
Department of Applied Computer Science
Vienna University of Economics and Business Administration
Augasse 2-6, A-1090 Vienna, Austria

REINHARD NECK
Department of Economics
Vienna University of Economics and Business Administration
Augasse 2-6, A-1090 Vienna, Austria

Abstract

In this paper we describe the algorithm OPTCON which has been developed for the optimal control of nonlinear stochastic models. It can be applied to obtain approximate numerical solutions of control problems where the objective function is quadratic and the dynamic system is nonlinear. In addition to the usual additive uncertainty, some or all of the parameters of the model may be stochastic variables. The optimal values of the control variables are computed in an iterative fashion. First, the time-invariant nonlinear system is linearized around a reference path and approximated by a time-varying linear system. Second, this new problem is solved by applying Bellman's principle of optimality. The resulting feedback equations are used to project expected optimal state and control variables. These projections then serve as a new reference path, and the two steps are repeated until convergence is reached. The algorithm has been implemented in the statistical programming system GAUSS.

1 Introduction

Many economic problems can be viewed as involving the optimization of an intertemporal objective function by a decision-maker who is constrained by a dynamic system subject to various kinds of uncertainties. Such problems arise both on the level of the individual firm as well as on the level of policy-making for a national economy. Stochastic optimum control theory has proved to provide a powerful methodology to deal with such problems. Stochastic optimum control problems are usually rather complex, thus for most models only numerical solutions can be obtained for particular values of the parameters. Even then, in most cases only approximations to the true optimum solution can be found at present. Thus, there is a need for further algorithmic developments. In the present paper we report on a new algorithm for the optimum control of nonlinear dynamic models that allows for additive uncertainty as well as for the presence of a stochastic parameter vector in the system equations, to be called OPTCON. In its present version, OPTCON is limited by two simplifications which prevent the solutions of turned to be truly optimal. First, computations of approximately optimal policies are obtained by applying repeated linearizations to the given nonlinear economic model. Second, we exclude any learning about the system parameters. A GAUSS implementation of the algorithm exists which can be obtained at request from the author mentioned first. More details on the algorithm and on some applications are given in [6].

2 The Optimum Control Problem

OPTCON can deliver approximate solutions to stochastic optimum control problems with a quadratic objective function and a nonlinear multivariable dynamic model in discrete time under additive and parameter uncertainties. Thus, we consider an intertemporal objective function which is additive in time and can be written as

$$L = \sum_{t=S}^{T} L_t(x_t, u_t) \tag{1}$$

with

$$L_t(x_t, u_t) = \frac{1}{2}\begin{pmatrix} x_t - \tilde{x}_t \\ u_t - \tilde{u}_t \end{pmatrix}' W_t \begin{pmatrix} x_t - \tilde{x}_t \\ u_t - \tilde{u}_t \end{pmatrix}. \tag{2}$$

x_t denotes an n-dimensional vector of state variables, u_t denotes an m-dimensional vector of control variables. The n-dimensional vector $\tilde{x}_t$ and the m dimensional vector $\tilde{u}_t$ denote the given "ideal" levels of the state and control variables, respectively. S denotes the initial and T the terminal period of the finite planning horizon.

The matrix W_t is defined as

$$W_t = \begin{pmatrix} W_t^{xx} & W_t^{xu} \\ W_t^{ux} & W_t^{uu} \end{pmatrix}, \qquad t = S, \ldots, T, \tag{3}$$

where W_t^{xx}, W_t^{xu}, W_t^{ux}, and W_t^{uu} are $(n \times n)$, $(n \times m)$, $(m \times n)$, and $(m \times m)$ matrices, respectively. Furthermore, we require W_t to be constant apart from a constant discount rate α:

$$W_t = \alpha^{t-1} W, \qquad t = S, \ldots, T, \tag{4}$$

where W is a constant matrix. Without loss of generality it is assumed that W is symmetric, which entails that

$$W^{xu} = [W^{ux}]'. \tag{5}$$

Defining

$$\begin{pmatrix} w_t^x \\ w_t^u \end{pmatrix} = -W_t \begin{pmatrix} \tilde{x}_t \\ \tilde{u}_t \end{pmatrix}, \tag{6}$$

$$w_t^c = \frac{1}{2}\begin{pmatrix} \tilde{x}_t \\ \tilde{u}_t \end{pmatrix}' W_t \begin{pmatrix} \tilde{x}_t \\ \tilde{u}_t \end{pmatrix}, \tag{7}$$

(2) can equivalently be written as

$$L_t(x_t, u_t) = \frac{1}{2}\begin{pmatrix} x_t \\ u_t \end{pmatrix}' W_t \begin{pmatrix} x_t \\ u_t \end{pmatrix} + \begin{pmatrix} x_t \\ u_t \end{pmatrix}' \begin{pmatrix} w_t^x \\ w_t^u \end{pmatrix} + w_t^c \tag{8}$$

The dynamic system is assumed to be given by the system of nonlinear difference equations

$$x_t = f(x_{t-1}, x_t, u_t, \theta, z_t) + \epsilon_t, \qquad t = S, \ldots, T, \tag{9}$$

where θ denotes a p-dimensional vector of unknown parameters, z_t denotes an l-dimensional vector of non controlled exogenous variables, and ϵ_t is an n-dimensional vector of additive disturbances. θ and ϵ_t, $t = S, \ldots, T$, are assumed to be independent random vectors with known expectations ($\bar{\theta}$ for θ, 0_n for ϵ_t, $t = S, \ldots, T$) and covariance matrices ($\Sigma^{\theta\theta}$ for θ, $\Sigma^{\epsilon\epsilon}$ for ϵ_t, $t = S, \ldots, T$). f is a vector-valued function.

*Financial support from the "Fonds zur Forderung der wissenschaftlichen Forschung" (research project no. P 5846) is gratefully acknowledged

3. Overview of OPTCON

Input of the Algorithm

system function	$\mathbf{f}(\ldots..)$
initial values of state variables	$\hat{\mathbf{x}}_{S-1} \equiv \mathbf{x}^*_{S-1}$
tentative path of control variables	$(\hat{\mathbf{u}}_t)^T_{t=S}$
path of non-controlled exogenous variables	$(\mathbf{z}_t)^T_{t=S}$
expected values of system parameters	$\hat{\theta}$
covariance matrix of system parameters	$\Sigma^{\theta\theta}$
covariance matrix of system noise	$\Sigma^{\epsilon\epsilon}$
weighting matrices of objective function	$\mathbf{W}^{xx}, \mathbf{W}^{ux}, \mathbf{W}^{uu}$
discount rate of objective function	α
"ideal" path for state variables	$(\bar{\mathbf{x}}_t)^T_{t=S}$
"ideal" path for control variables	$(\bar{\mathbf{u}}_t)^T_{t=S}$

Output of the Algorithm

expected optimal path of state variables	$(\mathbf{x}^*_t)^T_{t=S}$
expected optimal path of control variables	$(\mathbf{u}^*_t)^T_{t=S}$
expected optimal welfare loss	J^*_S

Description of the Algorithm

1. *Compute a tentative state path:* Use the Gauss-Seidel algorithm, the tentative policy path $(\hat{\mathbf{u}}_t)^T_{t=S}$, and the system equation $\mathbf{f}(\ldots..)$ to calculate the tentative state path $(\hat{\mathbf{x}}_t)^T_{t=S}$ according to

$$\hat{\mathbf{x}}_t = \mathbf{f}(\hat{\mathbf{x}}_{t-1}, \hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\theta}, \mathbf{z}_t) + \hat{\boldsymbol{\epsilon}}_t. \tag{10}$$

2. *Nonlinearity loop:* Repeat the steps (a) to (e) until convergence is reached (i.e., until the optimal control and state variables calculated do not change more than a prespecified small number from one iteration to the next) or the number of iterations is larger than a prespecified number. This procedure extends the one developed in [1].

(a) *Initialization for backward recursion.*

$$\mathbf{H}_{T+1} = \mathbf{0}_{n\times n}, \tag{11}$$
$$\mathbf{h}^x_{T+1} = \mathbf{0}_n, \tag{12}$$
$$h^\epsilon_{T+1} = 0, \tag{13}$$
$$\mathbf{h}^u_{T+1} = \mathbf{0}, \tag{14}$$
$$h^\rho_{T+1} = 0. \tag{15}$$

(b) *Backward recursion:*

Repeat the following steps i. to ix. for $t = T, \ldots, S$.

i. Compute the expected values of the parameters of the linearized system equation

$$\mathbf{x}_t \approx \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{B}_t\mathbf{u}_t + \mathbf{c}_t + \boldsymbol{\xi}_t, \qquad t = S, \ldots, T, \tag{16}$$

with

$$\mathbf{A}_t = (\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\mathbf{F}_{\mathbf{x}_{t-1}}, \tag{17}$$
$$\mathbf{B}_t = (\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\mathbf{F}_{\mathbf{u}_t}, \tag{18}$$
$$\mathbf{c}_t = \hat{\mathbf{x}}_t - \mathbf{A}_t\hat{\mathbf{x}}_{t-1} - \mathbf{B}_t\hat{\mathbf{u}}_t, \tag{19}$$
$$\boldsymbol{\xi}_t = (\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\boldsymbol{\epsilon}_t, \tag{20}$$
$$\Sigma^{\xi\xi}_t = \operatorname{cov}_{S-1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t) = (\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\,\Sigma^{\epsilon\epsilon}\left[(\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\right]', \tag{21}$$

where $\mathbf{I}_n$ denotes the n-identity matrix; $\mathbf{F}_{\mathbf{x}_{t-1}}, \mathbf{F}_{\mathbf{x}_t}, \mathbf{F}_{\mathbf{u}_t}$, and $\mathbf{F}_\theta$ are matrices of partial derivatives of $\mathbf{f}$ with respect to $\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{u}_t$, and θ, respectively. In (17)–(21), all derivatives are evaluated at the reference values $\hat{\mathbf{x}}_{t-1}, \hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\theta}, \mathbf{z}_t$, and $\hat{\boldsymbol{\epsilon}}_t = \mathbf{0}_n$.

ii. The matrices $\mathbf{A}_t, \mathbf{B}_t$ and the vectors $\mathbf{c}_t$ are functions of the random parameter vector θ and are, therefore, random themselves. Both matrices can be written as collections of their column vectors:

$$\mathbf{A}_t = (\mathbf{a}_{t,1} \;\ldots\; \mathbf{a}_{t,n}), \qquad t = S, \ldots, T, \tag{22}$$
$$\mathbf{B}_t = (\mathbf{b}_{t,1} \;\ldots\; \mathbf{b}_{t,m}), \qquad t = S, \ldots, T. \tag{23}$$

Each of these column vectors as well as $\mathbf{c}_t$ are functions of θ. These functions are approximated by linear functions.

$$\mathbf{a}_{t,i} = \mathbf{D}^{\mathbf{a}_{t,i}}\theta \quad \begin{matrix} i = 1, \ldots, n \\ t = S, \ldots, T, \end{matrix} \tag{24}$$
$$\mathbf{b}_{t,j} = \mathbf{D}^{\mathbf{b}_{t,j}}\theta \quad \begin{matrix} j = 1, \ldots, m \\ t = S, \ldots, T, \end{matrix} \tag{25}$$
$$\mathbf{c}_t = \mathbf{D}^{\mathbf{c}_t}\theta \qquad t = S, \ldots, T, \tag{26}$$

where $\mathbf{D}^{\mathbf{a}_{t,i}}$, $\mathbf{D}^{\mathbf{b}_{t,j}}$, and $\mathbf{D}^{\mathbf{c}_t}$ denote matrices of partial derivatives of the elements of (22), (23), and $\mathbf{c}_t$, respectively, with respect to θ.

We reshape these matrices as

$$\mathbf{D}^{\mathbf{A}_t} \equiv \left[\operatorname{vec}\left((\mathbf{D}^{\mathbf{a}_{t,1}})'\right), \ldots, \operatorname{vec}\left((\mathbf{D}^{\mathbf{a}_{t,n}})'\right)\right], \tag{27}$$
$$\mathbf{D}^{\mathbf{B}_t} \equiv \left[\operatorname{vec}\left(\left(\mathbf{D}^{\mathbf{b}_{t,1}}\right)'\right), \ldots, \operatorname{vec}\left(\left(\mathbf{D}^{\mathbf{b}_{t,m}}\right)'\right)\right], \tag{28}$$
$$\mathbf{d}^{\mathbf{c}_t} \equiv \left[\operatorname{vec}\left((\mathbf{D}^{\mathbf{c}_t})'\right)\right]. \tag{29}$$

iii. Compute the derivatives of the parameters of the linearized system with respect to θ:

$$\mathbf{D}^{\mathbf{A}_t} = \left[(\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1} \otimes \mathbf{I}_p\right]\left[\mathbf{F}_{\mathbf{x}_t,\theta}\mathbf{A}_t + \mathbf{F}_{\mathbf{x}_{t-1},\theta}\right], \tag{30}$$
$$\mathbf{D}^{\mathbf{B}_t} = \left[(\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1} \otimes \mathbf{I}_p\right]\left[\mathbf{F}_{\mathbf{x}_t,\theta}\mathbf{B}_t + \mathbf{F}_{\mathbf{u}_t,\theta}\right], \tag{31}$$
$$\mathbf{d}^{\mathbf{c}_t} = \operatorname{vec}\left[\left((\mathbf{I}_n - \mathbf{F}_{\mathbf{x}_t})^{-1}\mathbf{F}_\theta\right)'\right] - \mathbf{D}^{\mathbf{A}_t}\hat{\mathbf{x}}_{t-1} - \mathbf{D}^{\mathbf{B}_t}\hat{\mathbf{u}}_t, \tag{32}$$

where $\mathbf{F}_{\mathbf{x}_{t-1},\theta}$, $\mathbf{F}_{\mathbf{x}_t,\theta}$, and $\mathbf{F}_{\mathbf{u}_t,\theta}$ denote matrices of second-order partial derivatives of $\mathbf{f}$, introduced in a way analogous to [3], and all derivatives are evaluated at the same reference values as above.

iv. Compute the influence of the stochastic parameters: Compute all the matrices Υ^{AKA}_t, Υ^{BKA}_t, Υ^{BKB}_t, υ^{AKc}_t, υ^{BKc}_t, and υ^{Kc}_t, the cells of which are defined by

$$\left[\Upsilon^{AKA}_t\right]_{i,j} = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{a}_{t,j}}\Sigma^{\theta\theta}(\mathbf{D}^{\mathbf{a}_{t,i}})'\right] \quad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, n, \end{matrix} \tag{33}$$
$$\left[\Upsilon^{BKA}_t\right]_{i,j} = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{a}_{t,j}}\Sigma^{\theta\theta}\left(\mathbf{D}^{\mathbf{b}_{t,i}}\right)'\right] \quad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, n, \end{matrix} \tag{34}$$
$$\left[\Upsilon^{BKB}_t\right]_{i,j} = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{b}_{t,j}}\Sigma^{\theta\theta}\left(\mathbf{D}^{\mathbf{b}_{t,i}}\right)'\right] \quad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, m, \end{matrix} \tag{35}$$
$$\left[\upsilon^{AKc}_t\right]_i = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{c}_t}\Sigma^{\theta\theta}(\mathbf{D}^{\mathbf{a}_{t,i}})'\right] \; i = 1, \ldots, n, \tag{36}$$
$$\left[\upsilon^{BKc}_t\right]_i = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{c}_t}\Sigma^{\theta\theta}\left(\mathbf{D}^{\mathbf{b}_{t,i}}\right)'\right] \; i = 1, \ldots, m, \tag{37}$$
$$\left[\upsilon^{Kc}_t\right] = \operatorname{tr}\left[\mathbf{K}_t\mathbf{D}^{\mathbf{c}_t}\Sigma^{\theta\theta}(\mathbf{D}^{\mathbf{c}_t})'\right]. \tag{38}$$

v. Convert the objective function from "quadratic-tracking" to "general quadratic" format:

$$\mathbf{W}^{xx}_t = \alpha^{t-1}\mathbf{W}^{xx}, \tag{39}$$
$$\mathbf{W}^{ux}_t = \alpha^{t-1}\mathbf{W}^{ux}, \tag{40}$$
$$\mathbf{W}^{uu}_t = \alpha^{t-1}\mathbf{W}^{uu}, \tag{41}$$
$$\mathbf{w}^x_t = -\mathbf{W}^{xx}_t\bar{\mathbf{x}}_t - \mathbf{W}^{xu}_t\bar{\mathbf{u}}_t, \tag{42}$$
$$\mathbf{w}^u_t = -\mathbf{W}^{ux}_t\bar{\mathbf{x}}_t - \mathbf{W}^{uu}_t\bar{\mathbf{u}}_t, \tag{43}$$
$$w^c_t = \tfrac{1}{2}\bar{\mathbf{x}}'_t\mathbf{W}^{xx}_t\bar{\mathbf{x}}_t + \bar{\mathbf{u}}'_t\mathbf{W}^{ux}_t\bar{\mathbf{x}}_t + \tfrac{1}{2}\bar{\mathbf{u}}'_t\mathbf{W}^{uu}_t\bar{\mathbf{u}}_t. \tag{44}$$

vi. The key idea of our algorithm OPTCON is to use Bellman's principle of optimality:

$$J_t^*(x_{t-1}) = \min_{u_t} E_{t-1}\left(L_t(x_t,u_t) + J_{t+1}^*(x_t)\right), \quad (45)$$

where $J_t^*(x_{t-1})$ denotes the loss which is expected at the end of period $t-1$ for the remaining periods $t,\ldots,T$ if the optimal policy is implemented during these periods. $E_{t-1}(.)$ denotes conditional expectation; x_{k-1}, $k = S,\ldots t$, and u_{k-1}, $k = S+1,\ldots t$, are known at the time when we have to decide about u_t. It can be shown that $J_t^*(x_{t-1})$ can be expressed as a quadratic function of x_{t-1}:

$$J_t^*(x_{t-1}) = \frac{1}{2}x_{t-1}'H_t x_{t-1} + x_{t-1}'h_t^x + h_t^c + h_t^s + h_t^p \quad (46)$$

for all periods $t = S,\ldots,T+1$, where H_t, h_t^x, h_t^c, h_t^s, and h_t^p are defined below. Here we introduce the following simplifying assumptions:

A. Each occurence of $E_{t-1}(.)$ is substituted by $E_{S-1}(.)$, and each occurence of $\mathrm{cov}_{t-1}(..)$ is substituted by $\mathrm{cov}_{S-1}(..)$ for all $t = S+1,\ldots,T+1$. Thus, we rule out any learning about the parameters of the model.

B Although A_t, B_t and c_t are, in general, nonlinear functions of θ, we will compute their expected values by evaluating the equations (17), (18), and (19) at the reference values $\hat{x}_{t-1}$, $\hat{x}_t$, $\hat{u}_t$, $E_{S-1}(\theta)$, z_t, and $\hat{\varepsilon}_t = 0_n$, which were true only in case of linear functions.

vii. Compute the parameters of the function of expected accumulated loss:

$$K_t = W_t^{xx} + H_{t+1}, \quad (47)$$
$$k_t^x = w_t^x + h_{t+1}^x, \quad (48)$$
$$\Lambda_t^{xx} = \Upsilon_t^{AKA} + A_t'K_tA_t, \quad (49)$$
$$\Lambda_t^{xu} = (\Lambda_t^{ux})', \quad (50)$$
$$\Lambda_t^{ux} = \Upsilon_t^{BKA} + B_t'K_tA_t + W_t^{ux}A_t, \quad (51)$$
$$\Lambda_t^{uu} = \Upsilon_t^{BKB} + B_t'K_tB_t + 2B_t'W_t^{xu} + W_t^{uu}, \quad (52)$$
$$\lambda_t^x = v_t^{AKc} + A_t'K_tc_t + A_t'k_t^x, \quad (53)$$
$$\lambda_t^u = v_t^{BKc} + B_t'K_tc_t + B_t'k_t^x + W_t^{ux}c_t + w_t^u, \quad (54)$$
$$\lambda_t^s = \frac{1}{2}\mathrm{tr}\left[K_t\Sigma_t^{\xi\xi}\right] + h_{t+1}^s, \quad (55)$$
$$\lambda_t^p = \frac{1}{2}v_t^{cKc} + h_{t+1}^p, \quad (56)$$
$$\lambda_t^c = \frac{1}{2}c_t'K_tc_t + c_t'k_t^x + w_t^c + h_{t+1}^c. \quad (57)$$

viii Compute the parameters of the policy feedback rule:

$$G_t = -\left(\Lambda_t^{uu}\right)^{-1}\Lambda_t^{ux}, \quad (58)$$
$$g_t = -\left(\Lambda_t^{uu}\right)^{-1}\lambda_t^u. \quad (59)$$

ix. Compute the parameters of the function of minimal expected accumulated loss

$$H_t = \Lambda_t^{xx} - \Lambda_t^{xu}\left(\Lambda_t^{uu}\right)^{-1}\Lambda_t^{ux}, \quad (60)$$
$$h_t^x = \lambda_t^x - \Lambda_t^{xu}\left(\Lambda_t^{uu}\right)^{-1}\lambda_t^u, \quad (61)$$
$$h_t^c = \lambda_t^c - \frac{1}{2}\left(\lambda_t^u\right)'\left(\Lambda_t^{uu}\right)^{-1}\lambda_t^u, \quad (62)$$
$$h_t^s = \lambda_t^s, \quad (63)$$
$$h_t^p = \lambda_t^p. \quad (64)$$

(c) *Forward projection:*

Repeat the following steps i. and ii. for $t = S,\ldots,T$.

i *Compute the expected optimal policy:*

$$u_t^* = G_tx_{t-1}^* + g_t. \quad (65)$$

ii. *Compute the expected optimal state.* Use the Gauss-Seidel algorithm to compute x_t^* such that

$$x_t^* = f\left(x_{t-1}^*, x_t^*; u_t^*, \hat{\theta}, z_t\right). \quad (66)$$

(d) *Set the new tentative paths for the next iteration:*

$$(\hat{x}_t)_{t\le S}^T = (x_t^*)_{t\le S}^T, \quad (67)$$
$$(\hat{u}_t)_{t\le S}^T = (u_t^*)_{t\le S}^T. \quad (68)$$

(e) *Compute the expected welfare loss.*

$$J_S^* = x_{S-1}'H_Sx_{S-1} + x_{S-1}'h_{S-1}^x + h_{S-1}^c + h_{S-1}^p. \quad (69)$$

4 Concluding Remarks

In order to illustrate the feasibility of the algorithm OPTCON and its GAUSS implementation, we have applied it to a macroeconomic policy problem, the controls are fiscal and monetary policy variables, the states are macroeconomic target variables, and the objective function expresses a hypothetical policy-maker's preferences. We have used two small econometric models of the Austrian economy with eight (six) state variables, three control variables and three (two) non-controlled exogeneous variables. Different control experiments were performed, where the parameters of the model were regarded first as known with certainty, and then some stochastic parameters were tentatively introduced. Finally, the stochastic nature of all the parameters of the model (the coefficients and the constants) was taken into account by assuming the estimated values of the parameters to be their expected values and the covariance matrix of the coefficients of the model, which is delivered as an output of simultaneous estimations, to be the covariance matrix of the parameters. These optimum control experiments were carried out on an IBM compatible 12 MHz PC-AT with an 80287 mathematical co-processor. The running time of the GAUSS program of OPTCON ranged from 5 min. 5 sec. to 24 min. 37 sec. The results of the experiments show the influence of parameter uncertainty on the results of optimal policies for the two models considered.

Several directions of further research may be suggested. More optimization experiments are required in order to study the results under a greater variety of stochastic parameter patterns and different economic models. For the algorithm itself, there exist several possible extensions. By adding updating equations for the stochastic parameters such as the ones used in [4], it could be expanded into a passive-learning algorithm in the sense of [2]. Another interesting extension to be considered in the future will be the examination of the effects of decentralized policy-making, here results of decentralized control theory (dynamic team theory) and dynamic game theory will have to be incorporated.

References

[1] Gregory C. Chow, *Econometric Analysis by Control Methods* (Wiley, New York 1981).

[2] David Kendrick, *Stochastic Control for Economic Models* (McGraw-Hill, New York 1981).

[3] Elisabeth Chase MacRae, Matrix derivatives with an application to an adaptive linear decision problem, Annals of Statistics, 2(1974) 337.

[4] Elisabeth Chase MacRae, An adaptive learning rule for multiperiod decision problems, Econometrica 43(1975) 893.

[5] Josef Matulka and Reinhard Neck, OPTCON. An Algorithm for the optimal control of nonlinear stochastic models, Research Report 9007, Ludwig Boltzmann Institut für ökonomische Analysen wirtschaftspolitischer Aktivitäten, Vienna 1990.

HIERARCHICAL OPTIMAL CONTROL APPROACH FOR THE EUROPEAN COMMUNITY

Yukio Ito
Meijo University
Tenpaku, Nagoya 468
Japan
Aart de Zeeuw
Tilburg University
Postbox 90153
LE Tilburg, and
Free University,
Amsterdam, The Netherlands
Alejandra Cervio Pinho
UFSIA (SESO)
Antwerp, Belgium

Joseph Plasmans
UFSIA (Univ. Antwerp)
Prinsstraat 13
2000 Antwerpen
Belgium, and
Tilburg University
Tilburg, The Netherlands
Anton Markink
Tilburg University
Tilburg,
The Netherlands

Abstract - The purpose of this paper is to analyze the optimal coordination among interdependent economies with an independently acting coordinator, using aspects of hierarchical control theory in a linear (ized)-quadratic framework.

I-INTRODUCTION

Once the 1992 unification of Europe is achieved, the Single Market will reduce the room for independent policies as it might enlarge the existing imbalances among sectors, regions or, also, countries. To prevent this situation, the European Commission, or the Board of Prime Ministers, will have to act as a coordinator of more balanced economic policies.

II- THEORETICAL FRAMEWORK

In order to analyze this relationship under the assumption that Hierarchical Control Theory applies in this macroeconomic framework, we have built an annual interconnected econometric model, based on the COMET-V model (1988) for the Common Market. The countries under study are the EC countries less Luxemburg, and an aggregation of these countries (EC).

Trying to capture the interrelations that are taking place whithin this system, we can write its reduced-form as:

$$Y_i(t) = A_i\, Y_i(t-1) + A_{i0}\, Y_0(t-1) + B_i\, u_i(t) + D_i\, d_i(t) + \eta_i(t) \qquad (1)$$

where:

$Y_i(t)$: an n_i-dimensional vector containing the endogenous (target) variables of the i-th country, representing the objectives for the N subsystems (i=1,2,..,N) and the coordinator (i=0); A_i (i=0,1,...,N) is a matrix of the system parameters belonging to the lagged endogenous variables, and A_{i0} is the matrix of parameters belonging to the coordinator's target vector, denoting the effects of the coordinator's lagged endogenous variables on the i-th subsystem.

$u_i(t)$: a p_i-dimensional vector of the exogenous (control) variables which are subjected to control by the i-th policy maker and B_i is the matrix of coefficients of these control variables.

$d_i(t)$: a l_i-dimensional vector (called data-vector) including the current and lagged non-controllable exogenous variables of the i-th subsystem, and D_i is the matrix of coefficients of these pure exogenous data.

$\eta_i(t)$: an n_i-dimensional residual vector, assuming that

$E(\eta_i(t))=0$ and $E(\eta_i(t)\eta'_j(s))=\delta_{ij}\delta_{ts}\Omega_i$, where δ_{ij} and δ_{ts} are Kronecker deltas.

The Hierarchical Control Problem consists of the minimization of each subsystem's and coordinator's objective function, subjected to an econometric model. These objective functions are expressed by similar quadratic functions, which are the expected weighted sums of the squared deviations between actual targets and the desired targets, and the actual controls and the desired control variables. We assume that the subsystems would adopt a myopic behaviour, i.e., they would decide their policy period by period since they don't know the target variables of the coordinator and the other subsystems in advanced. The same holds for the coordinator.

$$J_i(t) = \min_{u_i(t)} E\left(\|Y_i(t) - Y^*_i(t)\|^2_{Q_i(t)} + \|u_i(t) - u^*_i(t)\|^2_{R_i(t)} \right) \qquad (2)$$

where Y^*_i and u^*_i, i=0,1,...,N, are the desired target and desired control variables, respectively. The weighting matrices $Q_i(t)$ (i=0,1,...,N) are assumed to be positive semi-definite and $R_i(t)$ (i=0,1,...,N) are assumed to be positive definite.

According to the Certainty Equivalence Principle, the optimal solutions will be the same in a stochastic optimal control problem with additive error term as in a deterministic optimal control problem. Taking the first derivatives of these objective functions with respect to the control variables, and equating these to zero, we obtain the control solutions.

III-APPLICATION ON A EUROPEAN MODEL

The model used is a linear(-ized) econometric model, with about 8 behavioural equations per country. These

equations refer to private expenditures (consumption and investment), foreign trade (total exports and imports), price indices (gross domestic product deflator and prices of total exports), unemployment rate and money stock. A definitional equation for trade balance was also introduced in the model. The interdependency among countries is measured by means of the quantity and price transmission in the international trade sector and by considering, explicitly, the lagged coordinator's values for unemployment rate, trade balance, gross domestic product and its deflator, in the countries' equations. The main control variables used in our model are the effective direct tax rates on household and company income, the indirect tax rate, and the rate of social security contribution of households. Other important control variables are monetary policy variables as the long or short term interest rate and the exchange rate with respect to the dollar (annual average), and the governmental subsidies, wages and employment. The model was estimated on a yearly basis for the sample period 1960-1986, and was applied to analyze the optimal coordination for the period 1987-1992.

In our experiment, the wheighting matrices were considered constant and diagonal. The Q_i-values concerning the GDP-deflator, unemployment rate, money stock and the deflator of total exports, were set equal to 2 and the other Q_i values were considered equal to 1. The R_i-values were set as an identity matrix.

The optimal control policies can be divided into different categories, depending on the direction of the information trans-mission. We concentrate our analysis on four type of informational exchange being defined as:

- Pure Top-Down Policy: the desired targets and desired controls determined by the coordinator, are transmitted into the subsystems' optimal policies.
- Pure Bottom-Up Policy: each subsystem determines its control policy, and transmits its information of targets to the coordinator, who determines his optimal control policy given the subsystems' targets.
- Mixed Top-Down Policy: the coordinator determines the overall target variables, but imposes no restrictions on the values of the subsystems' control variables.
- Mixed Bottom-Up Policy: the coordinator receives the information of targets from the subsystems, and transmits his desired control variables to the subsystems.

We compared the optimal targets and the optimal control variables with the observed values, by means of the Theil's inequality coefficient. This coefficient will measure the distance between the values obtained by the application of the hierarchical control solution and the observations, for which the OECD data (and its predictions until 1992) were taken.

The main conclusion is that there is not a significant difference neither between the Top-Down and the Mixed Top-Down policies, nor between the Bottom-Up and the Mixed Bottom-Up policies, so we may say that it is very important who determines the initial targets. Comparing among policies, and considering all the equations together, the Top-Down policy seems to be the more accurate for Germany, France, The Netherlands, Belgium, United Kingdom, Ireland, Denmark and Portugal. For the rest of the countries the Theil's values were similar.

For EC, Germany, France, The Netherlands and Belgium, the Theil's coefficients for the target variables were rather small, which can imply that the observed policy was close to the optimal. The rest of the countries present larger values, which implies that the observed policies are not considered to be the optimal.

Analyzing the Theil's coefficient for the control variables, the larger values were observed for the interest rate which may imply that a strong monetary policy is necessary. The government wages shows large values for Italy, The Netherlands, Ireland, Denmark and Portugal. For Irland, Denmark, Greece and Spain large values were also observed for the subsidies and the people employed by the government.

IV - CONCLUSION

This paper deals with the application of hierarchical control theory on a simultaneous econometric model for the European Community. By the analysis of the control solution we may conclude that for most of the countries a controlled policy seems to be near to the observed policy. However, analyzing equation by equation, it is difficult to determine which policy is "better" for the system as a whole. It will be fruitful to improve this policy approach by refining the econometric model. It will be also important to make some experiments variating the weighting matrices involved in the minimization problem, or setting different desired paths for the variables.

REFERENCES

- Ito, Y. and de Zeeuw, A. (1989). "Hierarchical Optimal Control Policy Approach in Econometric Models for the European Community: Integration or Decentralization". Working paper, Tilburg University.
- Findeisen, W. (1982). "Decentralized and Hierarchical Control under Consistency or Dissagreement of Interests", Automatica, vol.18, n° 6, pp. 647-664.
- Ito, y., Plasmans, J. and de Zeeuw, A. (1991). "Hierarchical Optimal Control Approach for Policy Evaluation in Econometric Models for the European Community". In preparation.

A UNIFYING TREATMENT OF REPAIR COST LIMIT MAINTENANCE POLICIES

Frank Beichelt
Ingenieurhochschule Mittweida
Department of Mathematics
O - 9250 Mittweida, Germany

Abstract - Repair cost limit maintenance policies are characterized as follows: When a system failure occurs the necessary repair cost is estimated. If the estimated cost exceeds a given limit, then the system is replaced by a new one; otherwise a minimal repair is carried out. Based on an inhomogeneous Poisson failure process a general and unifying approach to repair cost limit maintenance policies is presented. Important special cases are discussed in detail. Optimal repair cost limits are derived.

I. INTRODUCTION

During the past 40 years many mathematical models for proper scheduling of maintenance actions have been developed; for recent surveys see |4| and |9|. Their principal objective is to find maintenance policies being optimal with respect to given economic or reliability criteria. Among these policies the important class of repair cost limit replacement policies is characterized as follows: When a system failure occurs the necessary repair cost is estimated. If the estimated cost exceeds a fixed, possibly time-dependent limit - called repair cost limit - then the system is not repaired but replaced by a new one. This basic model may be combined with age and block replacement policies to obtain models being rather useful for practical applications. Usually the long-run average maintenance cost per unit time (maintenance cost rate) serves as an optimality criterion.

Repair limit replacement policies have been already proposed by Gardent and Nonant in 1963. Drinkwater and Hastings (1967) computed optimal repair cost limits by an approximation procedure which successively improves given repair cost limits. Under the same assumptions as stated by Drinkwater and Hastings (piecewise linear failure intensity, piecewise linear repair cost rate, exponentially distributed repair time) Beichelt (1978) derived the exact average cost per unit time so that optimal repair cost limits can be computed by any available optimization method. Later Park (1983) reconsidered a special case of this model. Based on a general failure model, in this contribution a unifying approach to repair cost limit models is developed allowing arbitrary repair cost distributions.

II. A GENERAL FAILURE MODEL

At time t=0 a system begins to work. Two types of system failures may occur:

Type 1: failures of this type are removed by minimal repairs.
Type 2: failures of this type removed by replacements.

By definition, a minimal repair does not affect the actual failure rate of the system |4|. Thus, type 1 failures may be interpreted as slight ones, easily to be removed, whereas type 2 failures may be complete system breakdowns. A failure occuring at system age t is with probability p(t) of type 2 and with probability $\bar{p}(t) = 1-p(t)$ of type 1. All replacement and minimal repair times are assumed to be negligible.

This failure model has been first introduced by Beichelt |1| in 1976, see |3| and |4| for improved versions. Later the special case $p(t)=p$ has been reconsidered by Fontenot and Proschan |6|. Here a short survey on those results is given needed for the purpose of this contribution.

The lifetime X of the system is assumed to be a random variable with IFR (increasing failure rate) distribution function F(t), survival function $\bar{F}(t) = 1-F(t)$, density function $f(t) = F'(t)$, and failure rate $q(t) = f(t)/\bar{F}(t)$. Let a cycle be the time between two neighbouring replacements and G(t) be the distribution function of the random cycle length Y, $\bar{G}(t) = 1-G(t)$. Then

$$P(t< Y \le t+\Delta t \mid Y> t) = p(t)q(t)\Delta t + o(\Delta t)$$

so that

$$\frac{G(t+\Delta t) - G(t)}{\Delta t}\Big/\bar{G}(t) = p(t)q(t) + \frac{o(\Delta t)}{\Delta t}.$$

Letting $\Delta t \to 0$,

$$G'(t)/\bar{G}(t) = p(t)q(t).$$

Hence p(t)q(t) is the failure rate belonging to G(t) so that

$$\bar{G}(t) = \exp\Big(-\int_0^t p(x)q(x)dx\Big). \qquad (1)$$

If X_k denotes the time of k th type 1 failure then it is well-known that on condition $Y>X_k$ the common probability density function of the random vector $(X_1,X_2,\dots,X_k)$ is |4|

$$f(x_1,x_2,\dots,x_k) = \begin{cases} q(x_1)q(x_2)\dots q(x_{k-1}), & x_1<x_2<\dots<x_k \\ 0, \text{ otherwise}, & k\ge 2. \end{cases} \qquad (2)$$

Let Z be the random number of type 1 failures within a cycle. Then,

$$P(Z=0) = \int_0^\infty p(t)dF(t), \text{ and for } n>1, \text{ using (2)}$$

$$P(Z=n) = \int_0^\infty \frac{1}{n!}\Big[\int_0^t p(x)q(x)dx\Big]^n p(t)dF(t).$$

Hence,

$$E(Z) = \sum_{n=0}^{\infty} nP(Z=n) = \int_0^\infty Q(t)dG(t)-1, \qquad (3)$$

where $Q(t) = \int_0^t q(x)dx$. In particular, for

$p(t) = p$,

$$E(Z) = p\int_0^\infty q(t)Q(t)e^{-pQ(t)}dt-1$$

$$= p\int_0^\infty x\, e^{-px}dx-1 = \frac{1-p}{p}\ ,\ p>0. \qquad (4)$$

If c_m and c_r denote the cost of a minimal repair and a replacement, respectively, then the long-run average maintenance cost under the failure model described is

$$K = \frac{\left[\int_0^\infty Q(t)dG(t)-1\right]c_m + c_r}{\int_0^\infty \bar{G}(t)dt}\ . \qquad (5)$$

In particular, for $p(t) = p$,

$$K = \frac{\frac{1-p}{p}c_m + c_r}{\int_0^\infty (\bar{F}(t))^p dt}\ . \qquad (6)$$

III. MAINTENANCE POLICY

Let C be the random cost for removing a system failure. Then the maintenance policy considered here is the following one:

On failure the system is replaced by an equivalent new one if C exceeds a given repair cost limit c; otherwise a minimal repair is carried out.

If c_r denotes the constant replacement cost and $R(x) = P(C\le x)$ the distribution function of C, then it is obvious to assume

$$R(x) = \begin{cases} 1, & \text{if } x\ge c_r, \\ 0, & \text{if } x<0, \end{cases} \quad \text{and } 0<c<c_r. \qquad (7)$$

The repair cost limit model can evidently be interpreted as a special case of the failure model discussed in section II: A type 1 (type 2) failure occurs if and only if $C\le c$ ($C>c$). Therefore,

$$\bar{p} = R(c),\ p = \bar{R}(c). \qquad (8)$$

In what follows if will be assumed that p does not depend on time, i.e. neither C nor c depend on the system age at failure. Hence (6) can be applied for computing the expected maintenance cost rate K=K(c). But an important peculiarity of the repair cost limit model has to be taken into account: the mean repair cost c_m for removing a type 1 failure depends now on c and, hence, on p:

$$c_m = \frac{1}{R(c)}\left[\int_0^c \bar{R}(x)dx - c\bar{R}(c)\right]. \qquad (9)$$

where $\bar{R}(x) = 1-R(x)$. (Note that c_m is the mathematical expectation of C on condition that $C\le c$.) Thus, (6), (8), and (9) yield the expected maintenance cost per unit time (maintenance cost rate)

$$K(c) = \frac{\frac{1}{\bar{R}(c)}\int_0^c \bar{R}(x)dx + c_r - c}{\int_0^\infty (\bar{F}(t))^{\bar{R}(c)}dt} \qquad (10)$$

The problem consists now in computing a repair cost limit $c=c^*$ minimizing (10). It will be solved on condition of Weibull-distributed lifetime: $\bar{F}(t) = \exp(-at^b)$. In this case,

$$K(c) = a^{1/b}\left[\Gamma(1+\tfrac{1}{b})\right]^{-1}\tilde{K}(c), \text{ where}$$

$$\tilde{K}(c) = (\bar{R}(c))^{1/b-1}\left[\int_0^c \bar{R}(x)dx+(c_r-c)\bar{R}(c)\right]. \qquad (11)$$

Hence the problem of minimizing K(c) is equivalent to minimizing $\tilde{K}(c)$. From $d\tilde{K}(c)=0$ there results a necessary condition for the optimal $c = c^*$:

$$\frac{1}{\bar{R}(c)}\int_0^c \bar{R}(x)dx + \frac{1}{b-1}c = \frac{1}{b-1}c_r\ . \qquad (12)$$

For $b>1$ there is a unique solution $c=c^*$ of (12). The corresponding minimum modified maintenance cost rate is

$$\tilde{K}(c^*) = \frac{b}{b-1}(c_r - c^*)(\bar{R}(c^*))^{1/b}\ .$$

Example 1. Let $R(x) = 1-\left[\frac{c_r-x}{c_r}\right]^s$; $0\le x\le c_r, s>0$.

Then the modified maintenance cost rate (11) is

$$\tilde{K}(c) = \left[\frac{c_r}{s+1}\left(\frac{c_r}{c_r-c}\right)^{s-s/b} + s\left(\frac{c_r-c}{c_r}\right)^{s/b+1}\right],$$

and the optimal repair cost limit is

$$c^* = \left(1 - \sqrt[s+1]{\tfrac{b-1}{b+s}}\right)c_r.$$

REFERENCES

[1] Beichelt, F.: A general preventive maintenance policy. Mathem. Operationsforschung und Statistik 7 (1976) 6, 927-932.

[2] Beichelt, F.: A new approach to repair limit replacement policies. Transact. of the 8th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, vol. C. Academia, Prague 1979, 31-37.

[3] Beichelt, F., Fischer, K.: General failure model applied to preventive maintenance policies. IEEE Trans. Reliab. R-29 (1980), 39-41.

[4] Beichelt, F.; Franken, P.: Zuverlässigkeit und Instandhaltung. Mathematische Methoden (Reliability and Maintenance. Mathematical Methods.) Carl Hanser Verlag München-Wien, 1984.

[5] Drinkwater, R.W.; Hastings, N.V.J.: An economic replacement model. Oper. Res. Quart. 18 (1967) 1, 59-71.

[6] Fontenot, R.; Proschan, F.: Some imperfect maintenance models. The Florida State University, Department of Statistics, FSU Statistical Repoart M 667, 1983.

[7] Gardent, P.; Nonant, L.: Entretien et renouvellement d'un parc de machines. Revue fr. Rech. Oper. 7 (1963) 1, 5-19.

[8] Park, K.S.: Cost limit replacement policy under minimal repair. Microelectronics and Reliability 23 (1983) 2, 347-349.

[9] Valdez-Flores, C.; Feldman, R.M.: A survey of preventive maintenance models for stochastically deteriorating single-unit systems. Naval Res. Log. Quart. 36(1989)4, 419-446.

ECONOMICALLY OPTIMAL PROCESS CONTROL USING SEQUENTIAL SAMPLING PLANS FOR VARIABLES

OLGIERD HRYNIEWICZ
Systems Research Institute
Newelska 6,Warsaw,POLAND

Abstract - Shewhart's $\bar{x}$-chart has been widely used for process control when quality characteristics are normally distributed. A general model for optimal design of control procedures has been proposed in Hryniewicz(1990).In the paper this model is used to design economically optimal sequential procedures.Comparison of the proposed optimal procedures and traditional Shewhart's control charts is presented.

1. INTRODUCTION

In 1924 Shewhart introduced a new method for controlling the quality of a production process-the control chart. Msthodology introduced by him consists of sampling from a process and evaluating the samples in order to find a signal that the considered production process is out-of-control.Whenever this state of the process is indicated searching and removing the assignable cause takes place. The problem of the optimal design of control charts using economic considerations was first solved by Duncan(1956).Since this very important work a great number of papers devoted to this problem have been published - especially to the problem of the optimal design of $\bar{x}$-charts. Original results by Duncan(1956) have been modified by von Collani(1986) who used as the objective function the average profit per item produced in the long run. The results of von Collani(1986) have been generalized in Hryniewicz(1990) for a very wide class of control procedures. In this paper we apply the results of Hryniewicz(1990) to develop economically optimal control procedures based on a sequential sampling plan for normally distributed variables.

2. MATHEMATICAL MODEL.

We assume, that the considered production process can be either in an acceptable STATE I characterized by a mean value μ_0 of normally distributed quality characteristic ,or in an unacceptable STATE II characterized by a shifted valuo $\mu_1=\mu_0+\sigma\delta$, where σ is a standard deviation of the controlled characteristic The values of μ_0, σ and δ are assumed to be known numerically.We further acsume that the duration of STATE I is a random variable distributed accordingly to a known continuous distribution function F(t).We assume that the moment of the transition from STATE I to STATE II is not directly observable and the transition from STATE II to STATE I can be achieved only by a special correction action.
We consider three types of possible actions:

a) monitoring of the process (sampling),
b) inspection (searching for an assignable cause),
c) renewal.

By monitoring we understand a statistical procedure which allows us to determine the actual state of the considered process with probability less than one.Thus,there are two types of error involved:type I error α, and type II error β.

We consider the following control procedure. After first h produced items we sample consecutive items observing their values x_i i=1,2,... . The test statistics has the following form

$$b + sn < \sum_{i=1}^{n} \left(\frac{x_i-\mu_0}{\sigma}\right) < a + sn \qquad (1)$$

where s=δ/2. We continue sampling until the two sided inequality (1) is fulfilled. When after n items observed the sum in (1) is greater than a+sn we stop the process and perform an inspection. When it is smaller than b+sn we stop our sampling procedure and do not interfere in the production process. The next sampling begins after next h-n produced items. The expected number of monitoring actions while the process operates in STATE I is given by (see e.g. Duncan(1956))

$$E[A_I]= \sum_{i=1}^{\infty} R(ih) \approx E[\tau]/h-0.5 \qquad (2)$$

where

$$R(ih)=1-F(ih) \qquad (3)$$

and E[τ] is the expected duration of STATE I. The expected number of monitoring actions while the process remains in STATE II is given by

$$E[A_{II}]= \frac{1}{1-\beta} \qquad (4)$$

Now we introduce economic parameters in order to describe the economic consequences of the actions to be performed. The cost of a sampling action is described by three quantities:

a_0^* - fixed cost per sampling,

a_1^* - unit cost of sampling,

n_i - the expected number of sampled elements,where i=1 in the case that the process is in STATE I during the sampling action, and i=2 in the case that it is in STATE II.

Moreover we assume that the following costs are known:

e^* - average cost of an erroneous inspection during STATE I,

r^* - average cost of searching for an existing assignable cause and the following renewal,

g_1 - average profit derived from one unit produced while the process is operating in STATE I,

g_2 - average profit derived from one unit produced while the process is operating in STATE II.

For the model described above it has been

shown in Hryniewicz(1988) that the optimal design of a control procedure for which the long term profit per unit produced is maximal can be found by maximization of the following objective function

$$G^*(h,\gamma)=e^*\left[\frac{b-E[A_I]\alpha-E[A_I]S_1-E[A_{II}]S_2}{h(E[A_I]+E[A_{II}])}\right]+g_2 \quad (5)$$

where

$$S_1=(a_0^*+a_1^*n_1)/e^* \quad (6)$$

$$S_2=(a_0^*+a_1^*n_2)/e^* \quad (7)$$

and

$$b=[(g_1-g_2)E[\tau]-r^*]/e^*>0 \quad (8)$$

It has been shown in Hryniewicz(1990) that the value of the optimal sampling interval approximately equals

$$h^*=E(\tau)\sqrt{\frac{2(\alpha+S_1)}{(2A_2-1)(b+A_2(\alpha+S_1-S_2)}} \quad (9)$$

where $A_1=E[A_I]$, and $A_2=E[A_{II}]$. The remaining parameters of the procedure have to be found by maximization of (5). When relative unit sampling costs a_1 are not too small approximately optimal sampling procedure can be found by minimization of (Hryniewicz(1990)):

$$G^{**}=(2A_2-1)(\alpha+S_1) \quad (10)$$

The approximately optimal values of parameters a and b obtained from minimization of (10) may visible differ from the values obtained by maximization of (5). Especially the approximately optimal parameter a is usually significantly greater than the optimal one. However, the resulting profits per unit produced differ in both cases only slightly.

3. PROPERTIES OF THE OPTIMAL PROCEDURES.

The main problem which has to be solved in the process of optimization is the calculation of α,β,n_1, and n_2 for the considered sequential sampling procedure. Unfortunately, there don't exist efficient algorithms for exact calculation of these parameters. Computer experiments have revealed that well known Wald's approximations cannot be applied. In our experiments we used approximations by Tallis and Vagolkhar(1965).

$$\alpha=1-L(-\delta/2) \quad (11)$$

$$\beta=L(\delta/2) \quad (12)$$

$$n_1=E(-\delta/2) \quad (13)$$

$$n_2=E(\delta/2) \quad (14)$$

where

$$L(\theta)=\frac{\Phi(-\theta)[1-\Phi(-\theta)-C(\theta)]}{D(\theta)(1-\Phi(-\theta))-\Phi(-\theta)C(\theta)} \quad (15)$$

$$C(\theta)=-(\exp(-2a\theta)/2)I(\theta) \quad (16)$$

$$D(\theta)=-(\exp(-2b\theta)/2)I(\theta) \quad (17)$$

$$I(\theta)=-\exp(\theta^2/2)(\theta+\theta^3/6+\theta^5/40+..) \quad (18)$$

$$E(\theta)=[(1-L(\theta)]a+L(\theta)b+$$
$$0.5[\Phi(-\theta)]^{-1}[1-L(\theta)][\phi(\theta)+\theta\Phi(\theta)]+$$
$$0.5[\Phi(-\theta)]^{-1}L(\theta)[\theta\Phi(-\theta)-\phi(\theta)]]/\theta \quad (19)$$

Numerous experiments have revealed that the objective function (5) is "flat" around its maximal value. Thus, different approximations lead to significantly different parameters of control procedures. However, in all those cases the value of the objective function remains practically the same.

To compare the optimal sequential procedure with traditionally designed $\bar{x}$-chart consider a production process with the following parameters:

$r^*=100.,e^*=50.,a_0^*=0.,g_1=1.,g_2=0.1,a_1^*=0.5$

Moreover, assume that the duration of STATE I is exponentially distributed with $\lambda=0,00005$. In Table 1 we compare the sequential procedure with $\bar{x}$-charts for 2 different values of shift δ. It is assumed that production rate is such that in the case of the $\bar{x}$-chart samples are taken after every 500 produced items.

TABLE 1
Comparison of optimal sequential procedures with traditional control charts

Parameters		δ=0.5	δ=2.
a		11.07	4.62
b	(seq.)	-0.6933	-0.3564
h		159	121
G		0.966619	0.986589
n		5	5
k	($\bar{x}$)	3.0	3.0
h		500	500
G		0.585848	0.977122

The results presented in Table 1, and similar results from many experiments reveal significant superiority of optimally designed sequential control procedures over $\bar{x}$-charts. The optimal sequential procedures are characterized by very small probabilities of false alarms α (usually smaller than 0.001) and small average sample sizes in STATE I (for moderate and high unit sampling costs usually close to 1.). Their superiority is especially significant for small shifts δ.

REFERENCES.

von Collani, E.(1986): A Simple Procedure to Determine the Economic Design of an $\bar{X}$-Control Chart, Journal of Quality Technology, Vol. 18, 145-151.

Duncan, A.J. (1956): The Economic Design of $\bar{X}$-Charts Used to Maintain Current Control of a Process, Journal of the American Statistical Association, Vol. 51, 228-242.

Hryniewicz, O. (1990): Approximately Optimal Economic Process Control for a General Class of Control Procedures. In: Frontiers in Quality Control IV (to be published)

Tallis, G.M., M.K. Vagholkar (1965): Formulae to Improve Wald's Approximation for Some Properties of Sequential Tests. JRSS ser.B, 74-81.

THE INSPECTION-ALLOCATION PROBLEM IN STATISTICAL PROCESS CONTROL CONSIDERED FROM AN ECONOMIC VIEWPOINT

BERNHARD F. ARNOLD
Fachhochschule München, Fachbereich 07
Lothstraße 34, W-8000 München 2, Germany

Abstract: Controlling a production line it is important to decide at which points of the line the process is to be inspected, i.e. which of the components are monitored simultaneously. Optimal control policies are defined by means of an economic objective function which can be interpreted as the average longterm profit per item produced. The method of dynamic programming turns out to be a useful tool for the determination of such optimal control policies. A numerical example is given in the case when several $\bar{x}$-charts are used in parallel.

I. THE PROCESS MODEL

We consider a production line consisting of m distinct and serially arranged components numbered 1,...,m, where production begins with component 1 and ends at component m. The number of items produced per time unit, i.e. the production rate $v>0$, is assumed to be constant in time and known. Furthermore, it is assumed that component j (j=1,...,m) is responsible for the value of the quality characteristic X_j, and we consider the case, where all the X_j's are mutually independent and univariate normally distributed.
The process starts in the in-control State I, the state of satisfactory production, where the expectations of the X_j's are equal to their target values μ_j, known:

State I: $X_j \sim N(\mu_j, \sigma_j^2)$; $j=1,\dots,m$; $\sigma_j^2>0$, known.

The time τ of production in State I is assumed to be exponentially distributed with expectation $1/\lambda>0$, known. When τ has passed a shock occurs and the process enters exactly one of m substate II(i) of the out-of-control State II; i=1,...,m:

State II(i): $X_j \sim N(\mu_j, \sigma_j^2)$; $j=1,\dots,m$; $j\neq i$;

$X_i \sim N(\mu_i \pm \delta_i\sigma_i, \sigma_i^2)$; $\delta_i>0$, known.

Letting C_i (i=1,...,m) denote the event of a transition from State I to State II(i), i.e. a failure of component i, then

$$q_i = \text{Prob}(C_i), \text{ known, with } \sum_{i=1}^{m} q_i = 1$$

is assumed to be constant in time.

Once the process produces in State II(i) we assume that it remains in this substate until a renewal actio is undertaken with respect to component i, after which the process starts anew in State I. The time between two subsequent renewals is called a renewal cycle.

II. THE CONTROL MODEL

Before performing a renewal action State II has to be detected. For this purpose we introduce a set of control policies, given by $((p;h(M_1),..,h(M_r);n(M_1),...,n(M_r);c_1,..,c_m))$.

Here p denotes any partition of the production line {1,2,...,m} into blocks $M_1,...,M_r$ $(1\le r\le m)$.

To give an example let be m=6 and p equal to 12|345|6. Then $M_1=\{1,2\}$, $M_2=\{3,4,5\}$ and $M_3=\{6\}$.

Furthermore,

$h(M_j)\in\mathbb{R}^+$ is the sampling interval at block M_j $(1\le j\le r)$,

$n(M_j)\in\mathbb{N}$ is the sample size used at block M_j $(1\le j\le r)$

and

$c_i\in\mathbb{R}_0^+$ is the control limit with respect to quality characteristic X_i $(1\le i\le m)$.

Such a control policy operates as follows: At the last component of each block M_j $(1\le j\le r)$ every $h(M_j)$ time units of production an independent random sample of size $n(M_j)$ ($n(M_j)$ consecutively produced items) is taken and analyzed with respect to all the quality characteristics X_i with $i\in M_j$: let $\bar{x}_i$ be the values of the corresponding sample means.

If $|\bar{x}_i-\mu_i|>c_i\sigma_i/\sqrt{n(M_j)}$ then component i — being an element of M_j — is inspected.

If $|\bar{x}_i-\mu_i|\le c_i\sigma_i/\sqrt{n(M_j)}$ holds for all $i\in M_j$, then M_j is left alone.

It is assumed that an inspection of component i $(1\le i\le m)$ reveals its actual state of production with probability one. If State II(i) is recognized a renewal is undertaken restoring State I.

Thus, a partition gives the components of the production line which are monitored simultaneously or — in other words — gives the points where monitoring and inspections are allocated. Furthermore, according to the here considered in-control and out-of-control states, two-sided $\bar{x}$-charts are used in parallel (one for each quality characteristic).

III. THE ECONOMIC MODEL

To determine the economic design of control charts Duncan (1956) introduces the average longterm profit per time unit as objective function. This model is also described by Montgomery (1985). Uhlmann (1982) and v.Collani (1981,1989) use the average longterm profit per item produced which turns out to be simpler from a mathematical point of view and seems to be more appropriate from the economic viewpoint, too. Thus, the author prefers the latter

approach: The here introduced objective function can be considered as a generalization of v. Collani's objective function. Let be P the total profit and N the total number of items produced, where both of the random variables refer to one renewal cycle. Then, the objective function is defined by

$$z^* = \sum_{i=1}^{m} q_i \frac{E[P|C_i]}{E[N|C_i]} .$$

To obtain an explicit expression of the objective function we introduce the following economic quantities:

$g_I > 0$: gain per item produced in State I.

$g_{II(i)} < g_I$: gain per item produced in State II(i).

$e_i > 0$: cost of an erroneous inspection of component i.

$a(M) > 0$: costs of sampling and analyzing one item with respect to all quality characteristics X_i with $i \in M$, where M is any block of the production line.

$b_i(M) > 0$: benefit of one renewal of component $i \in M$, where M is any block of the production line.

We assume that for any block M the costs of sampling n items and analyzing them with respect to all the X_i's with $i \in M$ are proportional to the sample size n, i.e. are equal to a(M)n.

In order to reduce the number of input-parameters, we maximize the standardized objective function

$$z := \left(z^* - \sum_{i=1}^{m} q_i g_{II(i)} \right)^{v/\lambda}$$

instead of z^* itself. With the standardized sampling intervals $x(M_j) := \lambda h(M_j)$ we get in analogy to Arnold (1990), where the case of r=1 is investigated:

$$z(p; x(M_1),\ldots,x(M_r); n(M_1),\ldots,n(M_r); c_1,\ldots,c_m) =$$

$$= \sum_{j=1}^{r} \frac{1}{x(M_j)} \left\{ - \sum_{i \in M_j} e_i \alpha_i - a(M_j) n(M_j) + \right.$$

$$\left. \sum_{i \in M_j} \frac{q_i (b_i(M_j)(1-\beta_i) + e_i \alpha_i)(\exp(x(M_j)) - 1)}{\exp(x(M_j)) - \beta_i} \right\} .$$

Here the α_i's and the β_i's are the probabilities of the errors of Type I and Type II respectively:

$$\alpha_i = 2\Phi(-c_i);$$

$$\beta_i = \Phi(c_i - \delta_i \sqrt{n(M_j)}) - \Phi(-c_i - \delta_i \sqrt{n(M_j)}),$$

where $i \in M_j$ and Φ is the standardized normal distribution function.

IV. THE SOLUTION TECHNIQUE AND AN EXAMPLE

Obviously the standardized objective function can be separated into r terms, where r is the number of blocks of the partition of the production line. These terms may be maximized separately and thus, the method of dynamic programming can be applied: Considering the production line $\{1,2,\ldots,i,i+1,\ldots,m\}$ and assuming that to all the "partial" production lines $\{m\},\{m-1,m\},\ldots,\{i+1,\ldots,m\}$ the optimal partitions are already calculated, then the optimal partition of the "partial" production line $\{i,\ldots,m\}$ is easily determined as the optimum of the partitions beginning with $i|\ldots|i,i+1|\ldots$ and ending at the block $\{1,2,..,m\}$, where $|\ldots$ indicates the optimal partitions of the "partial" production lines $\{i+1,..,m\}$ and $\{i+2,..,m\}$ respectively.

To give a numerical example we consider a production line consisting of six serially arranged components numbered $1,\ldots,6$. The probabilities q_i are assumed to be $q_1=0.1$, $q_2=q_3=q_4=q_5=0.2$, $q_6=0.1$. The numerical values of the shift parameters δ_i are $\delta_1=1$, $\delta_2=2$, $\delta_3=1.5$, $\delta_4=2$, $\delta_5=1$, $\delta_6=1.5$. The cost parameters e_i are $e_1=2$, $e_2=e_3=1$, $e_4=2$, $e_5=e_6=1$. The sampling costs per item are assumed to consist of the cost of 0.001 for taking one items and the cost of 0.0005 per quality characteristic for analyzing one item. For instance we obtain $a(\{3\})=0.0015$, $a(\{3,4\})=0.002$ and $a(\{2,3,4\})=0.0025$. Furthermore, let be $b_1=400$, $b_2=500$, $b_3=300$, $b_4=200$, $b_5=200$ and $b_6=300$. The benefit $b_i(M)$ with $i \in M$ is assumed to be equal to b_i diminished by 0.5 % per component between 1 and the sampling point determined by M. For instance we get $b_3(\{3\})=300$, $b_3(\{3,4\})=298.5$ and $b_3(\{3,4,5\})=297$. This approach of calculating $b_i(M)$ reflects the fact that it is desirable to detect failures in the production line as soon as possible. By dynamic programming and with $x(M)=\lambda h(M)$ we obtain the optimal control policy $(p; x(M_1),\ldots,x(M_r); n(M_1),\ldots,n(M_r); c_1,\ldots,c_6) = (1|2|34|56; 0.03519, 0.01160, 0.01564, 0.02563, 18,5,7,13; 3.102, 3.300, 3.118, 3.760, 2.870, 3.551)$ with $z=301.80$ as the corresponding maximum value of the standardized objective function. Thus, it is best to monitor the components 1 and 2 separately and to monitor simultaneously the components 3 and 4 on the one hand and the components 5 and 6 on the other hand.

References:

Arnold, B.F. (1990): An Economic $\bar{X}$-Chart Approach to the Joint Control of the Means of Independent Quality Characteristics. Zeitschrift fur Operations Research (ZOR) 34, 59-74.

Collani, E.v. (1981): Kostenoptimale Prüfpläne für die laufende Kontrolle eines normalverteilten Merkmals. Metrika 28, 211-236.

Collani, E.v. (1989): The Economic Design of Control Charts. Teubner-Verlag, Stuttgart.

Duncan, A.J. (1956): The Economic Design of $\bar{X}$-Charts Used to Maintain Current Control of a Process. Journal of the American Statistical Association (JASA) 51, 228-242.

Montgomery, D.C. (1985): Introduction to Statistical Quality Control. Wiley, New York.

Uhlmann, W. (1982): Statistische Qualitätskontrolle. Teubner-Verlag, Stuttgart.

STATISTICAL DESIGN OF FRACTION DEFECTIVE CONTROL CHARTS AND SOME ECONOMIC IMPLICATIONS*

Erwin M. Saniga
University of Delaware
College of Business and Economics
Newark DE 19716

Abstract

I develop an efficient computational algorithm for the statistical design of fraction defective charts when there are up to ten different shifts for which protection is desired. This algorithm allows the enumeration of all feasible statistical designs and thus can be extended to allow for the economic statistical design of fraction defective charts. A third use of this algorithm is to design single sample acceptance sampling plans with specified AQL and LTPD under a Type B scenario.

Statistical Design of Fraction Defective Control Charts With Some Extensions

I. Introduction

In this paper I develop an exact, computationally efficient algorithm to statistically design fraction defective charts. This algorithm can be used to give protection for up to ten shifts in the process parameter. A slight modification in input allows it to be used to design fraction defective charts for the situation in which small shifts in the process parameter are to be disregarded, a situation explored by Woodall(1985).

While this algorithm can be used on a stand alone basis, it can also be used as the basis for economic statistical design of fraction defective charts since it allows enumeration of all feasible statistical designs; these jointly define the feasible region for an economic statistical design.

In section II I discuss the development of the statistical design algorithm. Its use in its various forms for statistical design is illustrated with several examples. An example of its use in the design of a single sample fraction defective acceptance sampling plan with specified AQL and LTPD in a Type B situation is also illustrated.

In section III I use the algorithm to define the feasible region for economic statistical design of fraction defective control charts. A pattern search algorithm is developed to solve the economic statistical design problem and the results are compared to economic designs. A brief summary is drawn in section IV.

II. Statistical design

A statistical design of a fraction defective control chart involves the selection of the sample size n and the acceptance number c such that

$$ARL_1 > ARLB_1$$

and $$ARL_i < ARLB_i \qquad i=2,3,\ldots,k \qquad (1)$$

where ARL is the Average Run Length of the control chart when the process is in control with process proportion defective p_1 and ARL_i, i=2,3,..,k is the ARL when of shift of a specified magnitude p_i occurs ARLB is defined as the desired bound on ARL. An alternative to this type of statistical design is one in which it is desired that the control chart does not signal a shift when the shift is small. This type of design, which was proposed by Woodall can be defined as (1) above with the exception that for i=2, the constraint is replaced with

$$ARL_2 > ARLB_2.$$

We can define ARL in terms of α and β, respectively the Type I and Type II error probabilities of the control chart. Specifically,

$$ARL = 1/\alpha$$

and $$ARL_1 = 1/(1-\beta_1). \qquad (2)$$

For fraction defective control charts we can define α and β as

$$\alpha = 1-\sum_{j=0}^{c}\binom{n}{j}p_1^j(1-p_1)^{n-j}$$

$$\beta_i = \sum_{j=0}^{c}\binom{n}{j}p_i^j(1-p_i)^{n-j} \qquad i=2,3,\ldots,k \qquad (3)$$

Machine calculation of α and β can be difficult for practical values of n because of the factorial terms and powers of p in (3) above Recently, an efficient exact algorithm for this purpose was presented by Gehrlein, Ord and Fishburn(1986) and a similar algorithm is employed in my computations.

Statistical design of fraction defective charts can be accomplished by an enumeration of all n and c values until the constraints in (1) are all satisfied. A more efficient algorithm can be developed by using knowledge of the shape of the operating characteristic (O.C.) curves as n and c varies. Operating characteristic curves for various n and c are presented in Figure I. Specifically, as c increases for a fixed n the O.C. curve shifts to the right. As n increases for a fixed c the O C curve's steepness increases One can also summarize these changes in terms of of Type I and Type II error probabilities. In particular, as n increases α increases and β decreases for fixed c. As c increases α decreases and β increases for fixed n.

Using these facts I develop an algorithm more

efficient than enumeration. The steps in this algorithm are as follows. First, set c=0 and then check whether the first constraint is satisfied. If it is not then c can be increased until the first constraint is satisfied because any increase in n will further lower the first ARL. Next, n is increased until the remaining k-1 constraints are satisfied or until c is again increased.

I emphasize that there are a number of solutions to (1) and one must chose between these on the basis of some criteria. In the examples I solved, the criterion is the design satisfying (1) with the smallest n for the smallest c. The algorithm I develop allows enumeration of the feasible statistical designs within a finite range of n and c and an option in the coded version of the algorithm that is attached allows one to enumerate all feasible statistical designs.

III. Economic statistical design

In economic statistical design we wish to minimize the cost of a process or maximize the profit of a process when the costs of statistical process control and the costs of operating out of control are considered and statistical constraints are placed upon the model. Saniga(1989) developed the idea of economic statistical design and shows that this type of design can be more satisfactory for decision makers at all levels of a firm.

Mathematically, the economic statistical design problem can be expressed as

$$\begin{aligned} & \max\ P(n,c,h) \\ \text{s.t.}\quad & ARL_1 > ARLB_1 \\ & ARL_i < ARLB_i \qquad i=2,3,\ldots,k \end{aligned} \tag{5}$$

where P is the relative profit per item produced and h is the sampling frequency. In this paper, I use a model for P developed by Hryniewicz(1989) who explored the performance of economically designed fraction defective charts when shifts apart from expected shifts occured.

The function P is nonlinear and n and c are integer variables which complicates the solution procedure. In this paper, I take advantage of the statistical design algorithm I describe in the last section to facilitate the solution to (5). In particular, the algorithm I use to solve (5) is comprised of two stages. In the first stage, I use the statistical design algorthim to find a feasible statistical solution. Next, the optimal sampling frequency is found for this solution by pattern search. The same process continues for each feasible statistical solution. Finally, the feasible statistical solutions (with optimal h) are compared.

IV. Summary

I have designed an algorithm to allow the exact and efficient calculation of the design parameters for a fraction defective control chart under a statistical criterion. This algorithm, can also be used to design Type B single sample acceptance sampling plans with specified AQL and LTPD. The algorithm allows enumeration of the feasible region for the integer parameters in an economic statistical design and thus forms the basis for an efficient optimization algorithm for the economic statistical design. This use is demonstrated in the last section.

References

Chiu, W.K., (1975), "The Economic Design of Attribute Control Charts", Technometrics, 17, 81-87.

Deming, W.E., (1982), Quality, Productivity and Competitive Position, Cambridge, MA: MIT Press.

Duncan, A.J., (1974), Quality Control and Industrial Statistics, Irwin, Inc., Houewood, ILL.

Gehrlein, W., Ord, J.K., and P. Fishburn (1986), "The Limiting Distribution of a Measure of the Voting Power of Subgroups", Communications in Statistics, Vol. 15, No. 2, pp. 571-577.

Hryniewicz, O. (1989), "The Performance of Differently Designed P-Control Charts in the Presence of a Shift of Unexpected Size", Economic Quality Control, Vol. 4, No. 1, pp. 7-18.

Ishikawa, K (1976), Guide to Quality Control, Asian Productivity Organization, Tokyo, Japan.

Juran, J. Gryna, F. and Bingham, R, (1979) Quality Control Handbook, McGraw-Hill Pubs., N.Y., N.Y.

Saniga, E. (1989), "Economic Statistical Control Chart Designs with an Application to $\bar{X}$ and R charts", Technometrics, Vol. 31, No. 3, pp. 313-320.

Saniga, E., and Shirland, L., (1977), "Quality Control in Practice, a Survey", Quality Progress, 10, 30-33.

v. Collani, E. (1989), The Economic Design of Control Charts, Teubner Verlag, Stuttgart, Germany.

Woodall, W.H., (1985), "The Statistical Design of Quality Control Charts", The Statistician, 34, 155-160.

ECONOMICALLY OPTIMAL CONTROL OF PROCESS VARIABILITY

John Sheil,
Dept. Industrial Engineering,
University College,
Galway, Ireland.

Abstract. The derivation of economically optimal procedures to monitor and control production process variability is considered. The objective is to maximise average profit per item produced.

INTRODUCTION

Production processes are operated in order to generate profit. The level of profit is clearly a function of the quality of the items produced, which in turn depends on the operational state of the process. SPC procedures, typically $\bar{X}$,R charts and Cusum charts, are relatively widely used to control quantitative characteristics of produced items (and the process producing them). These procedures have traditionally been designed using statistical criteria only e.g.to satisfy specific ARL or OC-curve requirements: the costs and economic benefits associated with operating them being largely ignored. The idea of designing SPC procedures so as to optimise the economic performance of the process being monitored has attracted steady academic attention since the 1950s, but little practical implementation. Mathematical models containing too many parameters and complex optimization procedures are cited as prime reasons for this lack of acceptance[1]. Recent work, particularly by von Collani and others in Würzburg, shows that when the chosen objective is to maximise long-run profit per produced item, the resulting objective function is simple in formulation and allows easy determination of optimal control procedures. See, for example, the tabular procedure for designing optimal $\bar{X}$ charts in [2] and nomograms for $\bar{X}$,np and c charts in [3]. This paper shows how the objective function is obtained and used to generate economically optimal s charts for controlling process variability.

THE PROCESS MODEL

Let X be the quality characteristic of interest. Assume $X \sim N(\mu,\sigma^2)$. The process operates in one of two possible states: State I, the 'in-control' state with $\sigma = \sigma_0$ or State II, the 'out-of-control' state with $\sigma = \delta\sigma_0$, where $\delta>1$. The process starts in State I and can slip to State II as the result of a single assignable cause. The current state is determinable only by means of a process examination. A return from State II to State I requires a repair/renewal action. State I lifetime T, is exponentially distributed with mean $E(T) = 1/\theta$. The average number of items produced per hour is v and δ,v,θ,σ_0 are known.

THE CONTROL PROCEDURE

Samples of n items are drawn from current production at time intervals of length h. The standard deviation s is determined and plotted on a control chart with a single action limit at $k\sigma_0$. A process investigation, followed if necessary by a renewal, is undertaken if s plots outside this limit.
Clearly, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ (Chi-square distribution, with n-1 degrees-of-freedom). Hence, the probability of a 'false alarm' is given by

$$\alpha = \text{Prob}(s>k\sigma_0|\text{State I}) = 1 - F([n-1]k^2), \quad (1)$$

the probability of not detecting State II is

$$\beta = \text{Prob}(s\leq k\sigma_0|\text{State II}) = F([n-1]k^2/\delta^2) \quad (2)$$

and the average run lengths in State I, II are $ARL_I = 1/\alpha$, $ARL_{II} = 1/(1-\beta)$ respectively. F is the c.d.f. of a χ^2_{n-1} random variable. The parameters n,h,k are chosen so as to maximise the long-run average profit per item.

THE OBJECTIVE FUNCTION

Let a* = cost of sampling a single item,
e* = cost of investigating a false alarm,
b* = (expected) benefit per renewal viz. the extra profit resulting from a renewal/repair minus the cost of locating/repairing the assignable cause.
g = (average) profit per item when in State II

Further, for each renewal cycle, let A_I, A_{II} be the number of samples taken while the process is in State I, II ; let A_F be the number of false alarms and $A = A_I + A_{II}$ the total number of samples drawn. If L, N, G are the cycle length, number of items produced and the profit/gain achieved in a given cycle, then

$$E(N) = hE(L) = hvE(A), \quad (3)$$
$$E(G) = b^* + gE(N) - e^*E(A_F) - a^*nE(A). \quad (4)$$

Since T is exponential, it can be shown that

$$E(A_I) = 1/(e^{\theta h}-1) \text{ and } E(A_F) = \alpha/(e^{\theta h}-1). \quad (5)$$

Clearly,

$$E(A_{II}) = ARL_{II} = 1/(1-\beta). \quad (6)$$

Now, defining long-run average profit per item as $P^* = E(G)/E(N)$, substituting from equations (3)-(6) and re-writing, we obtain $P^*(h,n,k) =$

$$\frac{e^*}{vh}\left[\frac{(b^*/e^*)(e^{\theta h}-1) - \alpha}{e^{\theta h}-\beta}(1-\beta) - (a^*/e^*)n\right] + g$$

This is further simplified by noting that P^* is maximised with respect to h,n,k only by maximising $P(h,n,k)$

$$= \frac{1}{\theta h}\left[\frac{(b^*/e^*)(e^{\theta h}-1)-\alpha}{e^{\theta h}-\beta}(1-\beta) - (a^*/e^*)n\right]$$

Finally, setting $x = \theta h$ (viz. $h \div 1/\theta$), $b = b^*/e^*$ and $a = a^*/e^*$, we obtain the objective function:

$$P(x,n,k) = \frac{1}{x}\left[\frac{b(e^x-1)-a}{e^x-\beta}(1-\beta) - an\right], \quad (7)$$

where α, β are given by (1), (2) respectively. This objective function is very much simpler than earlier alternatives which were generally based on minimising the loss per unit of time, see [4], [5] for example. The parameters a,b are interpretable as the relative sampling cost per item and relative benefit per renewal, while x is the sampling interval expressed in average State I lifetime - a process characteristic. In practice, b will be large($\gg$1), while a and x will both be small - typically $\ll$1.

THE NO-SAMPLING PROCEDURE

When unit sampling cost(a^*) is high, relative to the cost of a process investigation(e^*) and the benefit(b^*), it is clear that from an economic viewpoint, it may be better not to sample/operate the control chart, but rather to perform routine process investigations (and if necessary, renewals) at regular intervals of length h. This procedure is described in terms of the s-chart parameter set (h,n,k) by setting n=0, k=0. In this case α=1; β=0 and the objective fonction(7) takes the form

$$P(x,n,k) = [b(e^x-1)-1]/(xe^x) \quad (8)$$

OPTIMAL CONTROL PROCEDURES

The task of finding the optimal s-chart design has been reduced to that of maximising P with respect to x,n,k (given that values for the economic parameters a^*, b^*, e^* are available). Despite the relatively simple appearance of (7), this is not a trivial task and will hardly be undertaken by the industrial practitioner. The report by von Collani and Sheil[6], contains extensive tables of optimal designs, indexed by values of δ,a and b: Table 1 contains a representative sample of these.

TABLE 1 : Some Optimal Designs

δ	b	a = .0005			a = .005			a = .05		
		n	k	x	n	k	x	n	k	x
1.5	20	36	1.333	.0387	19	1.306	.0930	0	.000	.3457
	50	37	1.330	.0247	19	1.309	.0569	0	.000	.2124
	150	37	1.330	.0142	19	1.311	.0321	0	.000	.1197
	500	37	1.330	.0077	19	1.312	.0174	0	.000	.0646
2.0	20	14	1.639	.0231	9	1.596	.0592	5	1.395	.1610
	50	14	1.639	.0145	9	1.599	.0367	5	1.412	.0958
	150	14	1.640	.0083	10	1.570	.0227	5	1.423	.0530
	500	14	1.640	.0046	10	1.571	.0124	5	1.430	.0283
2.5	20	9	1.882	.0184	6	1.844	.0467	4	1.603	.1344
	50	9	1.882	.0116	6	1.846	.0291	4	1.617	.0812
	150	9	1.882	.0067	7	1.78	.0189	4	1.626	.0454
	500	9	1.881	.0037	7	1.781	.0103	4	1.590	.0271
3.0	20	7	2.070	.0164	5	2.010	.0431	4	1.710	.1360
	50	7	2.071	.0103	5	2.013	.0269	4	1.722	.0829
	150	7	2.071	.0060	5	2.014	.0154	4	1.730	.0466
	500	7	2.073	.0032	5	2.015	.0084	4	1.735	.0251

A significant feature of the optimal designs, which is also apparent in Table 1, is that for each value for δ, the values for n and k are almost independent of b. This feature was used in [7] to provide tables of - at least approximately - optimal n,k pairs (and associated values for α,β) for a wide range of δ, a values. As anticipated, values for the 'standardised' sampling interval x are small. This fact leads to a simple method for computing a close approximation to x. Differentiating (7) with respect to x and equating to 0, yields the identity

$$(e^x-\beta+xe^x)(1-\beta)[b(1-\beta)+\alpha] = (e^x-\beta)^2[b(1-\beta)-an].$$

Taking the usual series expansion for e^x and neglecting $O(x^2)$ terms, yields the approximation

$$x \approx \left[\frac{2(1-\beta)^2(an+\alpha)}{(1+\beta)[b(1-\beta)+\alpha]}\right]^{\frac{1}{2}} \quad (9)$$

When the no-sampling procedure(n=0,k=0) is optimal, (9) becomes

$$x \approx \sqrt{[2/(b+1)]}. \quad (10)$$

In this case however, the exact value for x is easily obtained by iteration[7]. Thus, the problem of obtaining economically optimal s-charts to monitor process variability is greatly simplified for the potential user. The approximate approach outlined above - and detailed in [7] - has the disadvantage that the precise values for a,δ for which the control chart/procedure design is required may not appear in the table, so that some interpolation is required and a consequent decrease in accuracy likely results. The ideal would be to have nomograms available for the three design parameters; the author is currently working on the production of such nomograms.

REFERENCES

[1]. Montgomery, D.C.(1980). The Economic Design of Control Charts: A Review and Literature Survey, *Journal of Quality Technology*, 12, 75 - 87.

[2]. Collani, E.v.(1986). A Simple Procedure to Determine the Economic Design of an $\bar{X}$ Control Chart, *Journal of Quality Technology*, 18, 145-151.

[3]. Collani, E.v.(1989) The Economic Design of Control Charts, *Teubner-Verlag, Stuttgart*.

[4]. Duncan, A.J.(1956). The Economic Design of $\bar{X}$-Charts used to Maintain Current Control of a Process, *J.A.S.A.*, 51, 228-242

[5]. Lorenzen, T.J. and Vance, L.C.(1986). The Economic Design of Control Charts: A Unified Approach. *Technometrics*, 28,3-10.

[6]. Collani E.v. and Sheil, J.(1987). Economically Optimal s-Chart Designs. *Technical Report No.6, Würzburg Research Group on Quality Control, Uni.Würzburg, Germany.*

[7]. Collani E.v. and Sheil J.(1989). An Approach to Controlling Process Variability. *Journal of Quality Technology*, 21,87-96.

How Should the Four Parameters of a Sampling Plan Be Specified ? —— A Reasonable Specification of the Parameters from an Economic Viewpoint

Akihiro KANAGAWA and Hiroshi OHTA

Department of Industrial Engineering
College of Engineering
University of Osaka Prefecture
Sakai, Osaka 591, JAPAN

Abstaract - It is well known that the sampling plan with given producer's and consumer's risks is the most basic acceptance inspection. As it is, no sooner have we tried to use this plan than we are confronted by a difficulty, that is, how we should specify the four parameters determining the sampling plan. The four parameters are p_1, p_2, α and β, which specify the two points on the *OC*- curve. However, there are few studies on this topic. In this paper a reasonable specification of these parameters is discussed from an economic viewpoint.

1. INTRODUCTION

The sampling plan with given producer's and consumer's risks tends to be avoided to use because of the difficulty of specifying the two points on the *OC*- curve, that is (p_1, $1-\alpha$) and (p_2, β), where

- p_1 = fraction defective corresponding to an acceptable quality level,
- p_2 = fraction defective corresponding to an unsatisfactory quality level,
- α = specified value of producer's risk, and
- β = specified value of consumer's risk.

From this reason, Deming (1986) recommended to use the Dodge-Romig and MIL-STD sampling plans (ISO 2859). However, the sampling plan with given producer's and consumer's risks is most basic and useful if the difficulty of specifying the four parameters, p_1, p_2, α and β is dissolved. The purpose of this paper is to analyze the criterion of specifying the producer's and consumer's risks and to give a guidance for reasonable specification of these four parameters from an economic viewpoint.

2. INTERPRETATION OF SPECIFYING TWO-POINTS ON THE OC-CURVE

The criterion for specifying two points on the *OC* - curve, usually (p_1,$1-\alpha$) and (p_2, β), means that

1) To specify the probability that a good lot ($p = p_1$) is erroneously rejected is α.
2) To specify the probability that a nonconforming lot ($p = p_2$) is erroneously accepted is β.

Actual risks α^* and β^* are calculated as $1-L(p_1)$ and $L(p_2)$ respectively, where $L(p)$ is the probability of accepting a lot with fraction p defective.

From the producer's standpoint, the costs of accepting and rejecting a lot with fraction p defective without sampling inspection are

$$C_A^{(P)}(p) = NA_1^{(P)} + NpA_2^{(P)}$$
$$C_R^{(P)}(p) = NR_1^{(P)} + NpR_2^{(P)} \qquad (1)$$

where

N = lot size

$A_1^{(P)}$ = cost of accepting an item without regard to quality from the producer's standpoint

$R_1^{(P)}$ = cost of rejecting an item without regard to quality from the producer's standpoint

$A_2^{(P)}$ = cost of accepting a nonconforming item from the producer's standpoint

$R_2^{(P)}$ = cost of rejecting a nonconforming item from the producer's standpoint.

For details about a liner cost model, Hald's (1981) work can be referred to.

Hence, the following policy leads to a more profitable state :

If $C_A^{(P)} > C_R^{(P)}$, the lot is rejected (Case-$R_{(P)}$)

If $C_A^{(P)} < C_R^{(P)}$, the lot is accepted (Case-$A_{(P)}$).

When the lot is erroneously rejected although Case-$A_{(P)}$ occurs, the extra cost is

$$C_R^{(P)} - C_A^{(P)} = N(R_1^{(P)} - A_1^{(P)} + R_2^{(P)}p - A_2^{(P)}p.) \equiv N \cdot F(p). \qquad (2)$$

The expected value of this opportunity loss per lot, that is ' *Regret* ', is given as

$$Rg_1(p) = N \cdot F(p)(1 - L(p)). \qquad (3)$$

On the other hand, from the consumer's standpoint, the costs of accepting and rejecting a lot with fraction p defective without sampling inspection are

$$C_A^{(C)}(p) = NA_1^{(C)} + NpA_2^{(C)}$$
$$C_R^{(C)}(p) = NR_1^{(C)} + NpR_2^{(C)}, \qquad (4)$$

where the cost parameters from the consumer's standpoint correspond to those from the producer's standpoint.
In the same manner as the case for the producer, when the lot is erroneously accepted, the extra cost is

$$C_A^{(C)} - C_R^{(C)} = N(A_1^{(C)} - R_1^{(C)} + A_2^{(C)} - R_2^{(C)}) \equiv N \cdot G(p). \qquad (5)$$

The *Regret* in this case is

$$Rg_2(p) = N \cdot G(p)\, L(p). \qquad (6)$$

Thus the criterion for specifying the two points on the *OC*-curve (i.e. (p_1,$1-\alpha$) , (p_2 , β)) may be interpreted as a criterion having the following two regrets :

1) To specify the regret : $Rg_1(p_1) = N \cdot F(p_1)\,\alpha$ when a good lot ($p = p_1$) is erroneously rejected.
2) To specify the regret : $Rg_2(p_2) = N \cdot G(p_2)\,\beta$ when a nonconforming lot ($p = p_2$) is erroneously accepted.

From the above interpretation of the criterion, we may introduce the following constraints from the economic viewpoint :

$$Rg_1(p_1) \leq M^{(P)}$$
$$Rg_2(p_2) \leq M^{(C)}, \tag{7}$$

where $M^{(P)}$ and $M^{(C)}$ imply the allowable expected extra-loss for the producer and the consumer, respectively.

In the actual case that acceptance or rejection of lots is judged with sampling inspction, the inspection cost should be additionally considered. The inspection cost greatly depends on ASN (*Average Sample Number*).

Assuming that the actual inspection cost can be represented as

$$d_1 + d_2 s,$$

where d_1 = fixed cost for inspection, d_2 = sampling cost per item, and s = ASN. Then, in this case the respective regrets for the producer and the consumer are

$$Rg_1^*(p) = (N - s)\, F(p)\,(1 - L(p)) + \varepsilon\,(d_1 + d_2 s)$$
$$Rg_2^*(p) = (N - s)\, G(p)\, L(p) + (1-\varepsilon)(d_1 + d_2 s), \tag{8}$$

where ε denotes the proportion of sharing the inspection cost by the producer. From Eqs.(7) and (8), we have

$$Rg_1^*(p_1) \leq M^{(P)} + \varepsilon I$$
$$Rg_2^*(p_2) \leq M^{(C)} + (1 - \varepsilon) I, \tag{9}$$

where I = allowable inspection cost which is determined by the mutual agreement between the producer and the consumer. From Eqs.(3)(6) and (9), we have

$$1 - L(p_1) \leq \frac{M^{(P)} + \varepsilon(I - d_1 - d_2 s)}{(N-s)\, F(p_1)}$$
$$L(p_2) \leq \frac{M^{(C)} + (1-\varepsilon)(I - d_1 - d_2 s)}{(N-s)\, G(p_2)}. \tag{10}$$

From the definition of I, we have

$$I \geq d_1 + d_2 s, \tag{11}$$

then by substituting Eq.(11) into Eq.(10), we can determine the risks α and β as follows:

$$1 - L(p_1) \leq \frac{M^{(P)}}{(N-(1-d_1)/d_2)\, F(p_1)} \equiv \alpha$$
$$L(p_2) \leq \frac{M^{(C)}}{(N-(1-d_1)/d_2)\, G(p_2)} \equiv \beta. \tag{12}$$

Then, p_1 and p_2 are obtained by solving Eq.(12), that is

$$p_1 \equiv A_1^{(P)} - R_1^{(P)} + \frac{M^{(P)}}{(R_1^{(P)} - A_1^{(P)})(N-(1-d_1)/d_2)\,\alpha}$$
$$p_2 \equiv R_1^{(C)} - A_1^{(C)} + \frac{M^{(C)}}{(A_1^{(C)} - R_1^{(C)})(N-(1-d_1)/d_2)\,\beta}. \tag{13}$$

3. PROCEDURE FOR SPECIFYING FOUR PARAMETERS

It is said that the choice of (p_1, p_2) should be based on the technological requirements for usage of products and the economical consideration about the production cost, the inspection cost and so on. It is, however, difficult to specify (p_1, p_2) in actuality. Even though (p_1, p_2) can be specified, the specification of (α, β) is also difficult. As a matter of fact, most sampling inspection tables and design procedures are based on (α, β), which are some conventional values, e.g. 1%, 5%, 10%, etc. So the user of the sampling plan is obliged to specify (α, β) beforehand. By using the tentative (α, β) specified as the conventional values, we obtain (p_1, p_2) from Eq.(13), and then, a design of the single sampling attribute plan (n, c) is possible. Note that the *ASN* is derived after the sampling plan (n, c) was determined. Hence, it is necessary to ascertain whether Eqs.(10) and (11) are satisfied or not after determining the sampling plan (n, c).

Eq.(11) gives

$$1 - L(p_1) \leq \frac{M^{(P)} + \varepsilon(I - d_1 - d_2 s)}{(N-s)\, F(p_1)} \leq \alpha$$
$$L(p_2) \leq \frac{M^{(C)} + (1-\varepsilon)(I - d_1 - d_2 s)}{(N-s)\, G(p_2)} \leq \beta. \tag{14}$$

Then, if Eq.(11) is satisfied, Eq.(10) (equivalently, Eq.(9)) is satisfied. Even if Eq.(11) is not satisfied, it does not necessarily follow that Eq.(10) is not satisfied.

Table 1 shows the possible cases.

TABLE 1. Relationship Between Eqs.(10),(11) and User's Satisfaction With Their Requirements.

Case	Is Eq.(11) satisfied ?	Is Eq.(10) satisfied ?	Are the user's requirements satisfied ?
1	Yes	–	Yes
2	No	Yes	Yes
3	No	No	No

In the cases-1 or -2 in Table 1, the present sampling plan will be adopted. If the case-3 occures, it follows that

$$1 - L(p_1) \leq \alpha \leq \frac{M^{(P)} + \varepsilon(I - d_1 - d_2 s)}{(N-s)\, F(p_1)} \equiv \alpha'$$
$$L(p_2) \leq \beta \leq \frac{M^{(C)} + (1-\varepsilon)(I - d_1 - d_2 s)}{(N-s)\, G(p_2)} \equiv \beta' \tag{15}$$

Accordingly, there are following steps to get a sampling plan which satisfies Eq.(10), that is

Step-1 : To replace α and β with α' and β', respectively, and renew them α and β, respectively. The case-3 indicates that the initial specified risks were needlessly small. Sampling plan based on (α', β') is 'reduced' inspection in comparison with it based on old (α, β). So the new plan based on (α', β') will lessen its *ASN*.

Step-2 : To alter the type of inspection to that whose *ASN* can be reduced. For example, changing the sampling plan from *single* to *double*, or *double* to *sequential*. Enell(1984) discussed which sampling plan should be chosen in various situations.

Step-3 : To raise the allowable inspection cost I, through the mutual negotiation between the producer and the consumer.

References

Hald,A.(1981) *Statistical Theory of Sampling Inspection By Attributes*, Academic Press, London

Deming,W.E.(1986) *Out of Crisis* (Cambridge,MA, Massachusetts Institute of Technology)

Enell,J.W.(1984)"Which Sampling Plan Should I Choose ?", *J. of Qual. Tech.* Vol.16, No.3, pp.168-171.

The Adaptive Control Recursive Algorithm of Non-linear Economic System

Hao Zhigang Yang Yanmei
Institute of Applied Mathematics
Heilongjiang University, Harbin China

Abstract: In this paper, an adaptive control algorithm of non-linear economic system has been introduced. This algorithm is the dual of a parameter estimated algorithm, and it has the recursive form. The uniform small error of control can be obtained via this algorithm. Hence, it is satisfactory the application of economic system.

1. Introduction

There are a lot of economic systems, whose mathematic models are non-linear. For example, we consider the system which consists of M economic sectors. According to the Cobb-Douglas production function, we can write the mathematic model of the economic system as:

$$Q(t) = \sum_{i=1}^{M} A_i(t) L_i(t)^{a_i} K_i(t)^{b_i} \quad (1)$$

where t is a time, Q(t) is output, L_i is labour of ith sector, $K_i(t)$ is capital of the ith sector, $A_i(t)$ is "technical progress", a_i and b_i are elasticity of labour and capital of the ith sector, respectively.

It is very important for the analysis of the economic system above to solve the following problem: Suppose $Q_0(t)$ was given, how can define $L_i^0(t)$, $K_i^0(t)$, i = 1, 2,..., M, such that:

$$\sum_{i=1}^{M} A_i(t) L_i^0(t)^{a_i} K_i^0(t)^{b_i} = Q_0(t) \quad (2)$$

where we assume that $A_i(t)$, a_i and b_i, are known. This problem can be transfered into the problem of designing control system. The general form of this problem is:

Consider the non-linear economic system S. Suppose its model has been written in the form of prediction model:

$$y(t) = f[Y_{t-1}^{t-n}, U_t^{t-m}, \theta(t), t] \quad (3)$$

where y(t) is an one-dimensional output, u(t) is a input vector, $\theta(t)$ is an unknown random time-varying parameter vector, t is discrete time, and

$$Y_{t-1}^{t-n} = \{ y(t-n), y(t-n+1), \ldots\ldots, y(t-1) \}$$
$$U_t^{t-m} = \{ u(t-m), u(t-m+1), \ldots\ldots, u(t) \}$$

If the $y_0(t)$ was given, the problem is that how we can define $u^0(t)$, such that

$$f[Y_{t-1}^{t-n}, U_t^{t-m}, \hat{\theta}(t), t]$$

approximates enough to $y_0(t)$, where $\hat{\theta}(t)$ is an prediction value of $\theta(t)$.

The meaning of approximation is that the control algorithm satisfies the criterion of control error being uniform small.

Definition: Suppose $\{C\beta, \beta \in B\}$ is a group of control laws of the system S which contains parameter β. If for any $\varepsilon > 0$, there exist $\beta^0 \in B$ and $N > 0$, such that the control variable u(t) obtained with algorithm $C\beta^0$ for $t > N$ satisfies:

$$| y_0(t) - f[Y_{t-1}^{t-n}, U_t^{t-m}, \hat{\theta}(t), t] | < \varepsilon \quad a.s$$

where $y_0(t)$ is expected output at time t. $\hat{\theta}(t)$ is the prediction estimation of $\theta(t)$, we say that the control laws $\{C\beta, \beta \in B\}$ satisfy the criterion of control error being uniform small.

2 Adaptive control algorithm

For convenient, the model (3) of the system is changed into the following form:

$$y(t) = f[Y_{t-1}^{t-n}, u(t), U_{t-1}^{t-m}, \theta(t), t] \quad (4)$$

In general case, the adaptive control consists of two parts, which are the algorithm of parameter estimation and the algorithm of control law. But, in this paper, the parameter $\theta(t)$ of model (4) may be time-varying and the estimated value $\hat{\theta}(t)$ of $\theta(t)$ is necessary for defining u(t). Only can we obtain

$$\hat{\theta}(0), \hat{\theta}(1), \hat{\theta}(2), \ldots\ldots, \hat{\theta}(t-1),$$

via general estimated algorithm, so we must calculate prediction value $\hat{\theta}(t)$ of $\theta(t)$. This means that the

adaptive control algorithm must consist of three parts, which are the parameter estimated algorithm, the parameter prediction algorithm and the adaptive control law. More details are as following:

(1) The parameter estimated algorithm

In the paper [1] we have given estimated algorithm of parameter $\theta(t)$ in model (4):

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\delta}{\|\nabla_{\theta(t-1)}f[t,\hat{\theta}(t-1)]\|^2}\nabla_{\theta(t-1)}f[t,\hat{\theta}(t-1),t]\cdot[y(t)-f[Y_{t-1}^{t-n},u(t),U_{t-1}^{t-m},\hat{\theta}(t-1),t]] \quad (5)$$

where $\hat{\theta}(t)$ is an estimated value of $\theta(t)$, δ is a suitable constant and

$$\nabla_{\theta(t-1)}f[t,\hat{\theta}(t-1)]=\frac{\partial}{\partial\theta}f[Y_{t-1}^{t-n},u(t),U_{t-1}^{t-m},\theta,t]\Big|_{\theta=\hat{\theta}(t-1)}$$

The efficacy of this algorithm has been demonstrated by practical application and theoretical analysis [2].

(2) The parameter prediction algorithm

If t-1 is present time, using estimated algorithm (5), we only obtain estimative sequence

$$\hat{\theta}(0),\hat{\theta}(1),\ldots\ldots,\hat{\theta}(t-1)$$

This estimated value $\hat{\theta}(t)$ can only obtained by some prediction algorithm. Here, we suggest using multi-level recursive prediction method which details can be found in paper [3].

(3) The adaptive control law

Because the "position" of control variable $u(t)$ and unknown parameter $\theta(t)$ in model (4) may be considered as the same, as long as we consider the control variable $u(t)$ as time-varying parameter of the system, then we can determine control law $u(t)$ by estimated algorithm of unknown parameter: So we have

$$u^0(t)=u^0(t-1)+\frac{\beta}{\|\nabla_{u(t-1)}f[t,u^0(t-1)]\|^2}\nabla_{u(t-1)}f[u^0(t-1),t]\cdot[y_0(t)-f[Y_{t-1}^{t-n},u^0(t-1),U_{t-1}^{t-m},\hat{\theta}(t),t]]$$

where β is a suitable constant, $y_0(t)$ is expected output. $\hat{\theta}(t)$ is the prediction value of $\theta(t)$, and

$$\nabla_{u(t-1)}f[t,u^0(t-1)]=\frac{\partial}{\partial u}f[Y_{t-1}^{t},u,U_{t-1}^{t-m},\hat{\theta}(t),t]_{u=u^0(t-1)}$$

Under the proper condition, it can be prove that this adaptive control algorithm satisfies the criterion of control error being uniform small.

3 Application

Consider model (1) again, we suppose n=2, without using generality, then

$$Q(t)=\lambda_1(t)L_1(t)^{\alpha_{11}}K_1(t)^{\alpha_{21}}+\lambda_2(t)L_2(t)^{\alpha_{12}}K_2(t)^{\alpha_{22}}$$

Assume estimated values $\hat{\lambda}_1(t)$, $\hat{\lambda}_2(t)$, $\hat{\alpha}_{11}$, $\hat{\alpha}_{21}$, $\hat{\alpha}_{12}$ and $\hat{\alpha}_{22}$ were obtained. For the system, we have:

$$u(t)=(L_1(t),L_2(t),K_1(t),K_2(t))^T$$

Let:

$$d_1(t)=\hat{\alpha}_{11}\hat{\lambda}_1(t)L_1^0(t-1)^{\hat{\alpha}_{11}-1}K_1^0(t-1)^{\hat{\alpha}_{21}}$$
$$d_2(t)=\hat{\alpha}_{12}\hat{\lambda}_2(t)L_2^0(t-1)^{\hat{\alpha}_{12}-1}K_2^0(t-1)^{\hat{\alpha}_{22}}$$
$$d_3(t)=\hat{\alpha}_{21}\hat{\lambda}_1(t)L_1^0(t-1)^{\hat{\alpha}_{11}}K_1^0(t-1)^{\hat{\alpha}_{21}-1}$$
$$d_4(t)=\hat{\alpha}_{22}\hat{\lambda}_2(t)L_2^0(t-1)^{\hat{\alpha}_{12}}K_2^0(t-1)^{\hat{\alpha}_{22}-1}$$

then

$$\nabla_{u(t-1)}f[u^0(t-1),t]=(d_1(t),d_2(t),d_3(t),d_4(t))^T$$

If the expected output $Q_0(t)$ was given, we take that:

$$\begin{bmatrix}L_1^0(t)\\L_2^0(t)\\K_1^0(t)\\K_2^0(t)\end{bmatrix}=\begin{bmatrix}L_1^0(t-1)\\L_2^0(t-1)\\K_1^0(t-1)\\K_2^0(t-1)\end{bmatrix}+\frac{\beta}{\sum_{i=1}^{4}d_i(t)^2}\begin{bmatrix}d_1(t)\\d_2(t)\\d_3(t)\\d_4(t)\end{bmatrix}\cdot\{Q_0(t)-\hat{\lambda}_1(t)L_1^0(t-1)^{\hat{\alpha}_{11}}K_1^0(t-1)^{\hat{\alpha}_{21}}-\hat{\lambda}_2(t)L_2^0(t-1)^{\hat{\alpha}_{12}}K_2^0(t-1)^{\hat{\alpha}_{22}}\}$$

where $L_1^0(t-1)$, $L_2^0(t-1)$, $K_1^0(t-1)$ and $K_2^0(t-1)$ are labours and captials at time t-1 of two economic sectors respectively. For any $\varepsilon>0$, as long as we take the suitable value of β, for $\forall N$, we have

$$|Q_0(t)-\sum_{i=1}^{2}\hat{\lambda}_i(t)L_i^0(t)^{\hat{\alpha}_{1i}}K_i^0(t)^{\hat{\alpha}_{2i}}|<\varepsilon \quad a.s$$

Using this method, we can obtain a satisfied input of capital and labour in each economic sector.

Refrence

[1] Han Zhigang, On the Identification of Time-varying Parameter in Dynamic System, ACTA AUTOMATIC SINICA, VOL 10, NO. 4(1984), p330-337 (IN CHINESE).

[2] Han Zhigang, the Multi-level Recursive Method and Application, Science Press, Beijing (1989) (IN CHINESE).

[3] Han Zhigang, A New Method of Dynamic System Predication ACTA AUTOMATIC SINICA, Vol 9, No. 3 (1983), P161-168 (IN CHINESE).

STUDY ON MACROCONTROL MODEL OF TOTAL AMOUNT OF POLLUTANT FOR REGIONAL ENVIRONMENTAL POLLUTION

Zhang Baiqia
Chinese Research Academy of Environmental Sciences

Abstract-Development of economy, growth of population, exploitation of resource, progress of science and technology and environment are interdependent and mutually constrained so that there is a relationship of the unity of opposites among them and therefore, the study on environmental problem would have better understanding to control it only when the great system of unity of opposites is taken into accout. This study on the control model of total amount of pollutant in environment is just based on this viewpoint to establish an environmental econometric model for studying the pressure of economic development and population growth as well as the function of technical progress on environmental protection, analyzing and evaluating the status of environmental pollution, forecasting the developing tendency in future, and appraising interatively the different controlling alternatives to serve the decision making of environmental management.

1. Overall design of macrocontrol model of total amount of pollutant for regional environmental pollution

This macrocontrol model involves the integrative study on the generation, discharge, control or reduction and comprehensive utilization of the pollutants from economic sectors, resident's daily living and social consumption as well as the investment needed within the area of jurisdiction (province, autonomous region, municipality directly under the central government, city and prefecture). The overall structure designed as shown in fig.1

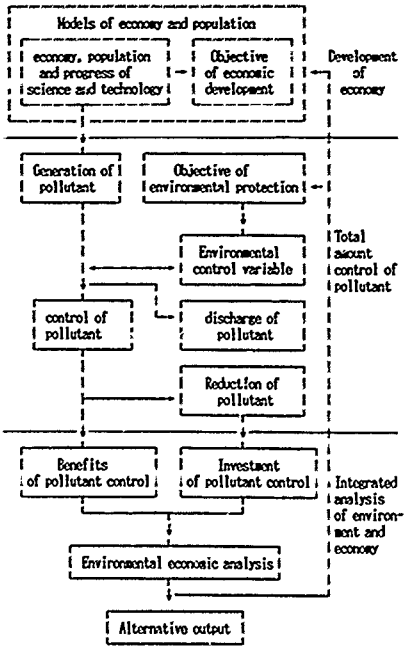


Fig.1 Overall structure of macrocontrol of total amount of pollutant

consists of three parts: the model of economy and population, the macrocontrol model of total amount of pollutant discharge in wastewater, waste gas and solid waste, and the model of integrated analysis of environment and economy.

A. Economy and population

Economic development and growth of population are the starting points of environmental pollution. The establishment of models of regional economic development and population control provides favorable condition not only for drawing up plan or program, forecasting and analyzing the situation of economic and social development, but also for studying the relation of environmental pollution to economic and social development. Under the condition with these models established, the models can be used in combination with the model of total amount control to get information about the impact of economic development and population growth on environment and also the constraint of environmental pollution on economic development, of which the quantitative analysis may provide the basis for decision making of the coordinative development of economy and environment. In case these models are not established, the indexes of economy and population may be used as external variables for the model of total amount control with the following requirements:

(1) Collection of information about the economy and urban population of the reference year according to the economic index of the "model".

(2) Collection of the similar information of the level year from related sectors in the area of jurisdiction.

The sectoral economy consists of two categories: that of province, autonomous region, large and medium cities is classified into 30 sectors according to the national statistic yearbook and the requirements of the statistic department of MEPA; and that of small cities lists all main pollution sources and controls more than 80% of pollutants in the area of jurisdiction. Based on the characteristics studied on the total amount of pollutants, the related indexes of corresponding economic sectors are selected as external variables for the model of total amount control to draw up planning alternatives with simulation of the reference year.

B. The macrocontrol model of total amount for waste water

This model is composed of two parts: the model of waste water from economic sector and the model of urban sewage. The pollutants are COD, oils and others selected according to the characteristics of the region. With the amount of generation of waste water and COD and also the output value of the sector in reference year, the generation coefficient of waste water per ten thousand dollar of output value and that of COD per ten waste water may be calculated using the model established. Considering the progress of science and technology, the coefficient of generation of pollutant in level year may be modified; and the amount of pollutant generated by the sector can be calculated with the output value in the level year (of the planning period or forecasting period). Using the control rate of waste water for three (high, medium and low) alternatives, the corresponding investment for the achievement of environmental objective of total amount control can then be derived through optimization. As to the urban sewage, the generation coefficient of domestic sewage per capita can be obtain similarly from the urban population and amount of sewage generated; and also the amount of sewage and COD based on specific conditions in the area of jurisdiction. Again, the control rate is used as variable to derive the investment for control and the benefits of sewage recycle may be obtained according to the amount of sewage recycled.

C. The macrocontrol model of total amount for atmospheric pollutants

This model contains three parts: (1) combustion process of economic sector generating waste gas, smoke and dust, SO2 and NOx; (2) technological process of economic sector generating waste gas, industrial dust, SO2 and NOx; and (3) fuel burning

process of resident's daily living and social consumption with pollutants same as (1), including the energy sources used by residents(such as coal, briquet, liquidized gas, natural gas and electric stove), centralized heat supply and the energy for enterprises and organizations. The amount of atmospheric pollutants generated in combustion process can be calculated on the basis of different fuel, type of combustion and discharge factor of pollutant; and that generated in technological process can be calculated similarly as for waste water, i.e. based on the generation coefficient of pollutant per ten thousand dollar of output value. The control variables for combustion process are the rates of smoke elimination and dedusting, etc.; those for technological process are control rate of waste gas, recovery rate of industrial dust and rate of rectilization of SO2, NOx and others; those for energy used in daily living are rate of gasification, percentage of briquet in the coal used (or the coefficient of energy structure for domestic usage);and that for centralized heat supply is the area of heat supply. With these variables mentioned above and the objective of total amount control of atmospheric pollutants, the amounts of various pollutants discharged, the amounts controlled, reduced and comprehensively utilized via various facilities and the corresponding investments of environmental protection can be calculated for three alternatives (high, medium and low). In order to achieve the macrocontrol objective with effective and rational investment of environmental protection, according to the results of calculation for the analysis and evaluation are carried out with criteria of value; all control variables are modified; and new alternative is established until it is satisfactory.

D.The macrocontrol model of total amount for solid waste

This model consists of two parts: the control model of industrial solid waste of economic sector and the control model of urban domestic refuse and manure. The industrial solid waste includes smelting residue,coal ash,slag,gangue,chemical residue, tailing and other residue; and its amount of generation is also derived from its generation coefficient per ten thousand dollar of output value, while that of coal ash or slag is calculated with the consumption of energy, that of gangue with its generation coefficient per ten thousand of raw coal, and that of domestic refuse and manure with their generation coefficient per capita of urban population in the area of jurisdiction. The control variables for industrial solid waste are the rate of comprehensive utilization and the rate of treatment and disposal; and those for domestic solid waste are the rate of mechanized cleanup and transport and the rate of detoxification. These variables are used to derive the amount of comprehensive utilization, treatment and disposal solid waste for three alternatives(high,medium and low), and also the corresponding investment of environmental protection. Similarly, the criteria of value is used to further modify the control variables to select rational investment for achieving the objective of environmental protection planning or forecasting period. The structure of macrocontrol model of total amount for waste water as show in Fig.2 demonstrates also the structure of the model for atmospheric pollutants and solid waste.

E.The integrated analysis of environment and economy

The model of integrated analysis is established on the basis of macrocontrol model of total amount for waste water, atmospheric pollutants and solid waste; and focused on the analysis and evaluation of: (1) the total investments for controlling various pollutants from different industries or enterprises and their ratio; (2) the total investment for controlling the pollutants from residents'daily living and social consumption; (3) the percentage of environmental protection investment of production sector to the investment of capital construction or renovation of the industies or enterprises of the sector; (4)the percentage of investment for environmental protection planning to the GNP in the planning period; and (5) the analysis of benefits of control,recovery and comprehensive utilization of various pollutants. In case modification is needed, users can base on the results of analysis and evalution to input new information to the control variables of related pollutant to get new solution through calculation until it is satisfactory.

II.function of technical progress in environmental protection-modification of generation coefficient of pollutant

the macrocontrol model of total amount of pollutant is dynamic with input of variables (such as the generation coefficient of pollutant) changing all the time because of continuous progress of science and technology and enhancement of management level. The change of generation coefficient of pollutant may influence directly the amount of pollutant generated, discharged or controlled as well as the investment for pollution control.

Hence, the modification of generation coeffcient of pollutant is the key to successful study of the macrocontrol model of total amount of pollutant, the change of which is related to following factors:

1. The technological renovation and the adoption of new technique, equipment and material in production sector increase the rate of utilization of energy and resource as well as economic benefits, thus lowering the generation coefficient of pollutant per ten thousand dollar of output value. In the years from 1984 to 1988 in China the discharge coefficient of industrial sewage was decreased by 9% each year; but the output value rapidly increased by 11.5% and the elasticity coefficient of the amount of sewage discharge was 0.2. This means technological progress has very great influence on environment.

2. Enhancement of comprehensive utilization of pollutant and development of technology generating less or no waste enable full utilzation of resource and decrease the generation coeffcient of pollutant.

3. Development of large-scale enterprise advanced in technology and more effective in scale may lower the consumption indexes of energy and resources in comparison with small-scale enterprise, because difference of the structure of large, medium and small enterprise, has marked influence on the generation coefficient of pollutant. For example, small cement plant generates more dust per unit output, about 6-8 times of that generated by medium cement plant.

4. Enhancement of management level of enterprise and implementation of the responsibility system for working post to reduce spattering or leaking and generation of pollutant in production process.

To sum up, the modefication of generation coeffcient of pollutant is of great significance.

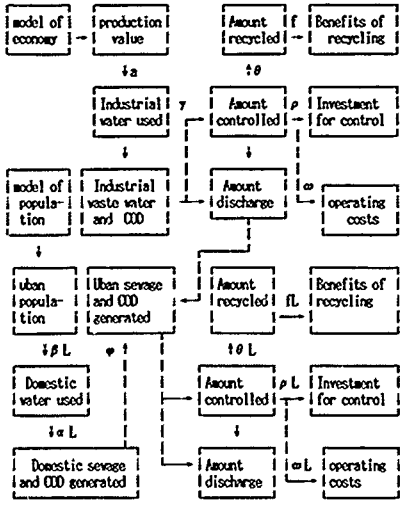


Fig.2 structure of macrocontrol model of total amount for waste water

Analysis of Customer Response to Time of Use (TOU) Rate in Ontario *

Sheldon X. C. Lou
Jieng Jiang
Feng Cheng
Faculty of Management
University of Toronto

C. W. Kenneth Keng
Economics and Forecasts Division
Ontario Hydro
Toronto, Ontario

Abstract

In this paper, we use pattern clustering and multivariate analysis of variance methods to analyze the daily load patterns in Ontario, Canada, in order to determine the impact of the Time Of Use rate implementation. Our findings show, that in the first year after the implementation, there were no significant changes in the electricity consumption behaviors of the customers selected for this study.

1 Introduction

Time-of-Use (TOU) rate has been proven to be effective in achieving desirable electricity load patterns [2], [3] In this paper, we analyze the DLPs in order to detect impact of the TOU rate on electricity demand patterns. Two approaches are employed The first one is the *pattern clustering* method, which partitions the monthly averaged DLPs for a number of years into different clusters (groups) according to some similarity measures. The resulting clusters are then analyzed to detect the potential pattern changes, and measure their significance. The second approach adopted is the well-known *multivariate analysis of variance* (MANOVA) method [1]. It compares the unified DLPs of different years in order to determine if they are statistically different.

The two methods are used to analyze the real electricity consumption data in Ontario, Canada.

2 Methodology

2.1 Pattern Clustering Method

The pattern clustering method proposed consists of three steps The first step is *data reduction and unification*. We take this step to compress the data, reduce random interference and remove the effects of economical or weather conditions. The second step is *clustering* The unified average DLPs produced by the first step are processed by the so-called *hierarchical clustering* method which partitions the DLPs into 2^ℓ clusters, $\ell = 1, 2, \cdots$. The clustering results are presented as tables of graphic patterns for visual inspection. The last step makes use of quantitative measures to assess the significance of the TOU impact.

2.1.1 Step 1: Data Reduction and Unification

For a particular year, denoting the DLP of day j, $j \in \{1, \cdots, 365\}$ as $d_j = [d_j^1\ d_j^2 \cdots d_j^{24}]$, then the monthly average $d_{mi} = [d_{mi}^1\ d_{mi}^2 \quad d_{mj}^{24}]$ of month i will be

$$d_{mi} = \frac{1}{m_{i+1} - m_i} \sum_{j=m_i}^{m_{i+1}-1} d_j,\ i = 1, \cdots, 12$$

where m_i is the number of days between January 1 and the first day of the i^{th} month plus 1.

We are only interested in DLP's shape but not its absolute value. So, we apply a unification approach to the monthly averaged d_{mi}. Then the unified d_{mi}, d'_{mi} is defined as

$$d'_{mi} = \frac{d_{mi}}{\bar{d}_i}$$

where

$$\bar{d}_i = \frac{1}{24} \sum_{k=1}^{24} d^k_{m_i}$$

Since we only use unified DLPs in the rest of this paper, the ' will be suppressed.

2.1.2 Step 2 : Clustering

The classical clustering method can be formally described as the following. Let $\mathcal{M}$ be an integer set, $\{d_j\}, j \in \mathcal{M}$ a set of objects $\{d_j\}$, and $n, 1 \le n \le M$ an integer where M is the number of objects in $\mathcal{M}$. The clustering method solves the following optimization problem :

$$J = \min_C \sum_{k=1}^{n} D(c_k) \qquad (1)$$

where $C = (c_1, c_2, \quad, c_n)$ is a partition of $\mathcal{M}$ with $c_1 \cup \cdots \cup c_n = \mathcal{M}$ and $c_i \cap c_j = \phi, i \neq j$.

$$D(c_k) = \min_{\bar{d}_k} \sum_{j \in c_k} \| d_j - \bar{d}_k \|^2 \qquad (2)$$

and $\| \cdot \|$ denotes the Euclidean distance Vectors $\bar{d}_k^*$ and partition C_* attaining the above minimum are called the cluster centroids and optimal partition respectively. It is not difficult to show, that $\bar{d}_k^* = \frac{1}{n_k} \sum_{j \in c_k} d_j$, where n_k is the number of objects in cluster k. To solve (1), several iterative algorithms are available [4]

In this paper, a variation of the classical clustering method called *hierarchical clustering* is used to generate 2^ℓ clusters It first divides all objects into two partitions (clusters) using Algorithm I These two partitions are called the first level partitions Then each partition at the first level is further divided into two smaller partitions called second level partitions, and so forth So, at level ℓ, there will be 2^ℓ partitions (clusters)

The hierarchical clustering results are presented as graphic pattern tables. An example in which four years' 48 average DLPs are analyzed is given in Fig 1. In this table, DLPs belonging to one cluster are represented by one graphic pattern By visually inspecting the table, one can easily observe three features. First, there are seasonal variations For each year, the average DLPs of May to September (summer season) belong to one cluster while that of October to April (winter season) belong to another at the second level. Second, years 86 to 88 seem to be very close to each other while year 89 looks very different. Third, the difference between 86 to 88 and 89 seems to be much larger than seasonal variations.

2.1.3 Step 4 : Using Quantitative Measures

We define the two distances :

1. *The Centroid Distance CD_{ij} is defined as the distance between the centroids of cluster i and cluster j*

*This research is partly supported by a grant from Ontario Hydro and a grant from NSERC.

2. *The Upper-Bound distance between clusters i and j is defined as the sum of the centroid distance and the objective function, and denoted by UD_{ij}. Thus $UD_{ij} = CD_{ij} + DBJ$.*

If we denote the maximum distance between the elements of two clusters as
$d_{ij} = \max_{a \in c_i, b \in c_j} \| d_a - d_b \|$, then we can show that CD_{ij} and UD_{ij} provides the upper and lower bounds of d_{ij}, i.e., $CD_{ij} \leq d_{ij} \leq UD_{ij}$.

We use the previous example to show that these two quantities can be used to measure the difference between two clusters.

We have mentioned that visual inspection indicated that the difference between 86-88 and 89 is a dominant factor. This statement can be quantified as follows. Let us turn to Fig. 2 which shows the distances between centroids and objective functions of the hierarchical clustering. Check the first level of Fig. 1 and 2. The difference between 86-88 (a cluster denoted by ◊) and 89 (♣) can be characterized by $CD_{\diamond\clubsuit} = 0.3189$ and $UD_{\diamond\clubsuit} = 0.3189 + 0.7336 = 1.0525$. The seasonal changes can be seen from the second level. For years 86-88, it is the distance between ◊ and ♡, and $CD_{\diamond\heartsuit} = 0.2224$ and $UD_{\diamond\heartsuit} = 0.2224 + 0.2261 = 0.4485$. For year 89, it is the distance between ♣ and ♠, and $CD_{\clubsuit\spadesuit} = 0.1237$ and $UD_{\clubsuit\spadesuit} = 0.1237 + 0.0234 = 0.1471$. Apparently, $CD_{\diamond\clubsuit}$ and $UD_{\diamond\clubsuit}$ are much larger than their counterparts $CD_{\diamond\heartsuit}$, $UD_{\diamond\heartsuit}$, $CD_{\clubsuit\spadesuit}$ and $UD_{\clubsuit\spadesuit}$. Thus, the difference between 86-88 and 89 is indeed bigger than seasonal changes.

Let us use the same example to show how the difference between years 88 and 89 is quantified. Since the difference between two years is defined as the average of the differences between their corresponding months, we first calculate the difference for each month. Let us start from January. At the third level, January 89 and January 88 belong to f_1(♣) and c_1(◊) respectively (see Fig. 2).

Since the approximation approach using two bounds cannot be directly calculated for f_1 and c_1, we go up 2 levels on Fig. 2 to the first level. Because January 89 is a member of a_2 and January 88 a member of a_1, we use the difference between a_1 and a_2 to approximate the distance between the two elements. We see from Fig. 2 that $CD_{a_1a_2} = 0.3189$ and $OBJ_{a_1a_2} = 0.7336$. Therefore $UD_{a_1a_2} = 0.3189 + 0.7336 = 1.0525$. Since February and March data are identical to January, we turn to April, and find that April 88 belongs to f_2(◊) and April 89 belongs to g_2(∞). Checking Fig. 2, we learn that the difference between c_1 and c_2 can be used to approximate that between f_2 and g_2. Simple computation shows that $CD_{c_1c_2} = 0.1237$ and $UD_{c_1c_2} = 0.1471$.

Repeating this process for the remaining months, we finally obtain

$$CD_{88,89} = \left(CD^{\text{Jan.}}_{88,89} + CD^{\text{Feb.}}_{88,89} + \cdots + CD^{\text{Dec}}_{88,89}\right)/12 = 0.286$$

and $UD_{88,89} = 0.902$.

Thus, roughly speaking, the difference between 88 and 89 lies between 0.286 and 0.902.

2.2 Multivariate Analysis of Variance Method (MANOVA) [1]

3 Experimental Results and Conclusions

Both pattern clustering and MANOVA methods were used to analyze the DPLs of the Ontario Western and Eastern Systems. None of these methods detected any statistically significant difference in the DLP's before and after the TOU rate implementation.

References

[1] G K Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*, New Yock. John Wiley, 1977.

[2] A Farugni, D J Aigner and R. T. Howard, *Customer response to time of use rate*, A report to the national Association of Regulatory Utility Commissioners, #84, 1981

[3] A Park, *Load controls and equipment for using off-peak energy*, A report to the National Association of Regulatory Utility Commissioners, #65, 1980.

[4] H Spath, *Cluster Dissection and Analysis*, New Yock. John Wiley & Sons, 1985.

n	year/mon	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2 first level	86	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊
	87	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊
	88	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊	◊
	89	♣	♣	♣	♣	♣	♣	♣	♣	♣	♣	♣	♣
4 second level	86	◊	◊	◊	◊	♡	♡	♡	♡	♡	△	◊	◊
	87	◊	◊	◊	△	♡	♡	♡	♡	♡	△	◊	◊
	88	◊	◊	◊	△	♡	♡	♡	♡	♡	△	◊	◊
	89	♣	♣	♣	♠	♠	♠	♠	♠	♠	♠	♣	♣
8 third level	86	□	◊	◊	□	△	♡	♡	♡	△	◊	□	□
	87	□	◊	◊	◊	△	♡	♡	♡	△	◊	□	□
	88	◊	◊	◊	◊	△	♡	♡	♡	△	◊	□	□
	89	♣	♣	♣	∞	∞	♠	♠	♠	∞	∞	♣	♣

Figure 1: Graphic Patterns for S1

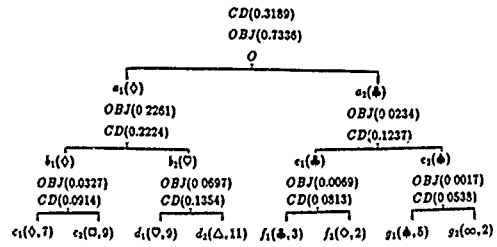


Figure 2: Hierarchical Clustering Structure for S1

Further Research on the Estimability of Linear Parametric Function in linear Models in Economic System

Zhou Shentang
Zhejiang Provincial Party School
Hang Zhou China

Abstract-The estimability of linear parametric functions in linear models have been studied by Baksalary and kala [1], Wang Songgui [2] and liu Shuangquan [3]. and some necessary and sufficient codtions for the estimibility of linear parametic functions have also been estibilished in this paper. we give another necessary and sufficient condition for the estimibility of linear parametric functions in liear model. And the result in this paper includes those which have been given in [2] and [3].

1. Introduction

For several ten years, linear models have been used in many fields of nation economy in our country. such as in industry, agriculture, meteorolgy, economy management, medical and health work, education and psychology, e.c. And extended the study of ther theories. In order to use the linear model to slove practical problems, it is nessesary to give a further study of the estimibility of linear parametric function in linear models.

A general linear model is:

$$Y = X\beta + e,\ E(e) = 0,\ \mathrm{cov}(e) = \sigma^2 V \tag{1}$$

Where Y is an $n\times 1$ vector of observations, X is an $n\times p$ known matrix, β is a $p\times 1$ vector of unkown parameters. and e is an $n\times 1$ random vector which with expectation zero and covariance $\sigma^2 V$, where V is a nonnegative definite matrix.

IF in model (1) β is restricted as $H\beta=b$, then the model is writen as follows:

$$Y = X\beta + e,\ H\beta = b,\ E(e) = 0,\ \mathrm{cov}(e) = \sigma^2 V \tag{2}$$

Where H is an $l\times p$ known matrix, b is an $l\times 1$ known vector, the others are the same as those in (1).

We know that the parametric functions $A\beta$ of β is estimable if there is an $m\times n$ matics B, such that $E(BY)=A\beta$. For all β, where A is an $m\times p$ matrix, We denote $\hat{\beta}$ as the least square estimator of β, then for estimable function $A\beta$, we have $A\hat{\beta}$ is the best linear unbiased estimator of $A\beta$ from the Gauss-Markov thorem when $|V|\neq 0$ where $|v|$ denote the determinant of the matrix V. In paractice, not all the linear functions of β are estimable. So some necessary and sufficient conditions for the estimability are given previously. Searle (1966) states the condition as: the linear combinations $A\beta$ are estimable if and only if $A(X'X)^-X'X=A$ where $(X'X)^-$ is any matrix satisfying the matrix equation $X'X(X'X)^-X'X=X'X$. (The x^- is called the g-inverse of Matrix X). Baksalary and Kala (1976) [1] state the condition as: $r(X(Ip-A^-A))=r(X)-r(A)$, where $r(X)$ denote the rank of the estimability of an matrix. Wang Songgui (1981) [2] give another condition of the estimability of $A\beta$ and avoiding the computation of g-inverse of a matrix. What to do in his paper is to slove linear homogenous equation and compute the rank of a matrix. Liu Shuangquan (1988)[3] also give another condition which include that in [2]. We now also give a condition which include those in [2] and [3]. So the condition in this paper is the extension of those in previous papers.

2. Preliminary

The sequel discuss is in real spaces. Let A,X,C,T,V be matrix and x,y, α,β,γ, be vectors.

Definition: Let $\alpha_1,\alpha_2,\cdots,\alpha_s$ be s linear independent vectors, then $L(\alpha_1,\alpha_2,\cdots,\alpha_s)$ is the space spaned by the vectors $\alpha_1,\alpha_2,\cdots,\alpha_s$. that is for every $\beta\in L(\alpha_1,\alpha_2,\cdots,\alpha_s)$, there exit $t_1,t_2,\cdots,t_s$; such that;

$$\beta = t_1\alpha_1 + t_2\alpha_2 + \cdots + t_s\alpha_s,$$

Lemma 1: Let A be $m\times p$ matix R^p,R^m be p-dimensional and m-dimensional spaces.

$A:R^p\to R^m$ be projector. then

$$1.\ \dim\ (R^p) = \dim(R(A)) + \dim(N(A)) \tag{3}$$

$$2.\ R^{\perp}(A) = N(A') \tag{4}$$

Where $\dim(\cdot)$ denote the dimension of a space,

$$R(A) = \{Ax; x\in R^p\}$$
$$N(A) = \{x; Ax = 0, x\in R^p\}$$
$$R^{\perp}(A) = \{x; x\perp R(A)\}$$

A' is the transpose of matrix A. Form lemma 1. We have:

$$p = \dim(R^p) = \dim(R(A)) + \dim(N(A)).$$

So the dimendion of solution of the equation $Ax=0$ is:

$$\dim(N(A)) = p - \dim(R(A)) = p - r(A) \tag{5}$$

Lemma 2: Let A,B be $m\times p$ and $p\times q$ matrices, then

$$r(AB) = r(B) - \dim(R(B)\cap N(A)) \tag{6}$$
$$= r(A) - \dim(R(A')\cap N(B')) \tag{7}$$

3. Main result

From [2] and [3] we have

Theorem 1: In linear model (1), $A\beta$ is estimable if and only if

$$r(A) + r(XV) = r(X) \tag{8}$$

Where A is $m\times p$ matrix, $r(A)=k$, V be $p\times(p-k)$ matix, Whose row vectors are the solution of $A'AY=0$.

Theorem 2: Let T be $p\times q$ matrix such that $N(A)=R(T)$. Then $A\beta$ is estimable in linear model (1) if and only if

$$r(A) + r(XT) = r(X) \tag{9}$$

From theorem 1 and theorem 2 we known that theorem 2 is the extension of theorem 1 since when for a $p\times q$ matix T such that $N(A)=R(T)$ theorem 1 is also hold. But there is no answer in [3] about the search of T. In this paper we give another condition which extended the matrix C, and also give the way of how to search C.

Theorem 3: Let C be $p\times q$ matrix, such that

$$\dim(R((X')\cap N(C')) = \dim(R(X')\cap R(A'))$$

Then $A\beta$ is estimable in linear model (1) if and only if

$$r(A) + r(XC) = r(X) \tag{10}$$

Proof: From lemma 2, we have

$$r(XC) = r(X) - \dim(R(X')\cap N(C'))$$

So $$r(XC) = r(X) - \dim(R(X')\cap R(A'))$$

From the definition of estimability, we know that $A\beta$ is estimable if and only if there esist a $m\times n$ matrix B such that $A=BX$, where $\mu(x)$ denote the space spaned by column of X. But $\mu(A')\subset\mu(X')$ is equivalent to:

$$\dim(R(x')\cap R(A')) = r(A)$$

So $$r(XC) = r(X) - \dim(R(X')\cap R(A')) = r(X) - r(A)$$
$$r(A) + r(XC) = r(X)$$

Note 1: From theorem 2, we know that T is used as C in theorem 3 but the acceptable spectrum of C is much greater than that of T. Beacause N(A)=R(T) state that N(A) and R(T) have the same maximum linear independent vectors, but C is different from that, this may be seen as follows:

Let $\alpha_1,\alpha_2,\cdots,\alpha_s,\alpha_{s+1},\cdots,\alpha_{t_1}$ be maximum linear independent vector of R(X'), where $t_1=r(X')=r(X)>s$; $\alpha_1,\alpha_2,\cdots,\alpha_s,\beta_{s+1},\cdots,\beta_{t_2}$ be maximum linear independent vectors of R(A'), where $t_2=r(A')=r(A)>s$; $\alpha_1,\alpha_2,\cdots,\alpha_s$, $\gamma_{s+1},\cdots,\gamma_r$ be maximum linear independent vectors of N(C') where $r=p-r(C')=p-r(C)>s$, and also: α_i, β_j, $\gamma_t(i=1,2,\cdots,t_1;\ j=s+1,s+2,\cdots,t_2;\ t=s+1,s+2,\cdots,r)$ are each other independent vectors. From the suppose above, we have

$$\dim(R(X')\cap N(C'))=s$$
$$=\dim(R(X')\cap R(A'))$$

So if there is also $r(XC)=r(X)-r(A)$.

We have that AB is estimable from theorem 3. But under this circumstance, $N(A)\neq R(C)$, we can not use theorem 2 to justify if AB is estimable or not. Even if $r(XC)=r(X)-r(A)$ is hold. When $N(A)=R(C)$ theorem 3 is obviously hold since from lemma 1, we have

$$N(C')=R^{\perp}(C)=N^{\perp}(A)=R(A')$$
$$\dim(R(X')\cap N(C'))=\dim(R(X')\cap R(A'))$$

So the condition $N(A)=R(C)$ is only a particular situation of

$$\dim(R(X')\cap N(C'))=\dim(R(X')\cap R(A')$$

So the results in our paper is the extension of those in [2] and [3]

Note 2: How to search for C?

Since $X_{n\times p}$, $A_{m\times p}$ are also known matrics, so R(X'), R(A') are also known

$$\text{Let } r(X')=t_1\leqslant\min(p,n)$$
$$r(A')=t_2\leqslant\min(p,m)$$

The maximum linear independent ventors of R(X'), R(A') are the same as those supposed in Note 1 Since

$$0\leqslant s\leqslant\min(t_1,t_2)$$

Then $\dim(R(X')\cap R(A'))=s$

Form theorem 3 if there is also

$$\dim(R(X')\cap N(C'))=s$$

Then the estimability of AB in linear model (1) can be justified. For if C has be searched for, we can do is only to justify if

$$r(XC)=r(X)-r(A) \text{ is true or not}$$

The ways of chosing N(C') are at least C_p^s. Beacause we can just chose s vectors from the maximum linear independent vectors of R(X'), then add any vectors which are independent with $\alpha_1,\alpha_2,\cdots,\alpha_s$: the vectors we chosen and the added vectors form a new linear indenpendent vectors whose dimension is smaller than the whole space considered.

Suppose the chosen vectors are $\alpha_{i_1},\alpha_{i_2},\alpha_{i_3},\cdots,\alpha_{i_s}$.

Where $i_1,i_2,i_3,\cdots,i_s$ is a permutation of $1,2,\cdots,t_1$ let $\gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r}$ be the adding vectors which are independedt with $\alpha_{i_1},\cdots,\alpha_{i_s}$ then $\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s}$, γ_{t_1}, $\gamma_{t_2},\cdots,\gamma_{t_r}$ form a new space as:

$L_0(\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r})$

Let $N(C')=L_0(\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r})$

And $M'=(\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r})$

Where M' is a matrix formed by $\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r}$.

From $N(C')=\{t_i=0,\ t\in R^s\}$

We have $C'(\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r})=0$

$C'M'=0$

So $MC=0$

And C can be get from sloving the eqution.

Note 3: If we chose $\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s}$ as $\alpha_1,\alpha_2,\cdots,\alpha_s,\gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r}$ as $\beta_{s+1},\beta_{s+2},\cdots,\beta_{t_2}$

then

$$M'=(\alpha_{i_1},\alpha_{i_2},\cdots,\alpha_{i_s},\ \gamma_{t_1},\gamma_{t_2},\cdots,\gamma_{t_r})$$
$$=(\alpha_1,\alpha_2,\cdots,\alpha_s,\beta_{s+1},\beta_{s+2},\cdots,\beta_{t_2})$$

Where $u=t_2-s$

So $M=A$

And $MC=0$ is equivalent to $AC=0$. Since $A'Ax=0$ have the same solution as $Ax=0$, then we get theorem 1.

$\{\text{the solution of } MC=0\}\supset\{t: Mt=0,\ t\in R^s\}\bigcap\{Ct, t\in R^t\}$

So the theorem 2 is a particular situation of theorem 3. For Linear model (2) we have

$$\begin{pmatrix}y\\b\end{pmatrix}=\begin{pmatrix}X\\H\end{pmatrix}\beta+\begin{pmatrix}e\\o\end{pmatrix},\quad E\begin{pmatrix}e\\o\end{pmatrix}=0,\quad Cov\begin{pmatrix}e\\o\end{pmatrix}=\sigma^2\begin{pmatrix}v&o\\o&o\end{pmatrix}$$

Then we have:

Corllary 1: Let C be a $p\times q$ matrix, such that

$$\dim(R(X'H')\cap N(C'))=\dim(R(X'H')\cap R(A'))$$

then AB in model (2) is estimable if and only, if

$$r\left(\begin{pmatrix}X\\H\end{pmatrix}C\right)=r\begin{pmatrix}X\\H\end{pmatrix}-r(A)$$

Where

$$(X'H')=\begin{pmatrix}X\\H\end{pmatrix}'$$

Let a multivariate linear model is

$$Y=X\beta+e \qquad (11)$$

Where Y is an $n\times m$ random matrix, X is an $n\times p$ known matrix, and β is a $p\times m$ unknown parameters, e is an $n\times m$ random matrix, let tr(X) denote the trace of matrix X, tr(AB) is estimable if there exist an $m\times n$ matrix B such that:

$$E(tr(By))=tr(AB)$$

From the definiton, we known that tr(AB) is estimable if and only if there exist a $m\times n$ matrix B such that:

$$A=BX$$

Corollary 2: Linear model (11) if there exist a $p\times q$ matrix C, such that

$$\dim(R(X')\cap N(C'))=\dim(R(X')\cap R(A'))$$

then tr(AB) is estimable if an only if

$$r(XC)=r(X)-r(A)$$

Reference

[1] J. K. Baksalary and R. Kala: extensions of Milliken's Estimability Crteria Ann. Statist. 4(1976) 639-642

[2] Wang Songgui: on the Estimebility of Linear Parametric Functions in Linear Models

A√TA Mathmatical Applicate SINICA Vol4 NO4 NOV 1981

[3] Liu Shuang quan Further Research on the Estimability of Linear Parmaetric Functions in Linear Models

Mathematical Statisivs and Applied Probability Vol 3 No2, 1988 165-168.

Dynamic Prediction Model of Regional Industry Economy System

ZHang Yujing
Statistics bureau of Harbin City

Liu Chunsheng
Statistics bureau of Heilong jiang province

Han Zhigang
Hei Long Jiang University

Abstract—In this paper, the Double-Level Prediction Model is raised and by using it, a new method to solve prediction problem of regional gross output value of industry is presented. The results of practical checking computation are satisfied.

I. Introduction

It is significance both in theory and practice to analyse and predict regional industry economy system. It is not only helpful to make middle-term or long-term plan scientifcly, but also has important guiding value in managing and adjusting, controlling the recent industry economy. Formerly general Time Series Analysis Method was mostly used to analyse and predict regional industry economy developing state. But the regional gross industrial output value isn't a Stationary Random Sequence, therefore, the application of classic Time Series analysis method will come across some difficulties. Sometimes the time series data have to be resolved. The tendency term and period term have to be separated. In case the other terms were stationary, the classic Time Series Analysis Method was used to predict the gross industrial output value.

By using the Double-level prediction Model and Multi-Level Recursive Prediction Method, It won't be necessary to resolve the time series and the supposition of the steady time series also isn't necessary. And the compution amounts of prediction will be decreased greatly.

II. Double-Level Prediction Model and Multi-Level Recursive Prediction Method

Thinking over the changing condition of gross industrial output value of one area, Let $Y(k)$ stand for the gross output value of industry at k time, let $U(k)$ stand for the input of industry system, including investment in fixed assets and circulating funds. Generally speaking, the changing law of $Y(k)$ can be described by the model of the form below.

$$Y(k)+a_1Y(k-1)+\cdots+a_nY(k-n)=b_0U(k-p)+b_1U(k-p-1)+\cdots+b_mU(k-p-m)+V(k) \quad (1)$$

where $a_1,\cdots a_n$, $b_0,\cdots b_m$ are awaiting parameters, $V(k)$ is Random Noise term, interference, n and m are ranks of the model, and p is Time Lag. The parameters $a_1,\cdots a_n$, $b_0,\cdots b_m$ of the model are Time Varying. The changes of them show some relevant changes of economy inner system. And the literature (1) indicates that the Random Noise $V(k)$ of model (1) can be omitted. But at this time, it must be compensated by the Parameter Randonilazation. So we can change model (1) into the form below

$$Y(k)+a_1(k)Y(k-1)+\cdots+a_n(k)Y(k-n)=b_0(k)U(k-p)+b_1(k)U(k-p-1)+\cdots+b_m(k)U(k-p-m) \quad (2)$$

among them, $a_1(k),\cdots a_n(k), b_0(k),\cdots b_m(k)$ are Random Time Varying parameters. n and m are ranks of the model p is Time Lag. They have certain economy meaning.

If $\varphi(k)=\{-Y(k-1),\cdots,-Y(k-n),U(k-p),U(k-p-1);\cdots U(k-p-m)\}^t$

$\theta(k)=\{a_1(k),\cdots a_n(k),b_0(k),\cdots b_m(k)\}^t$

then model(2) can be transformed into

$$Y(k)=\varphi(k)^t\theta(k) \quad (3)$$

where $\theta(k)$ is a multi-dimension random series. Model (2) or Model (3) is called First-Level Prediction Model.

Since $\theta(k)$ is a multi-dimension variable, the mathematics model $M(\Theta(k))$ that $\theta(k)$ can be satisfied can be created according to itself changing character. At this place,

$$\Theta(k)=\{\theta(k),\theta(k-1),\cdots,\theta(k-q)\}$$

here q is a proper constant. Since $\theta(k)$ hasn't had any demisions, model $M(\Theta(k))$ can be set up by the pure mathematics method. According to literature(2), the Model AR method, Constant Increment method, Constant Factor method, Period Variable method , Period Increment method, Average approximate method and etc. are usually used to set up model $M(\Theta(k))$ on prediction problem. Model $M(\Theta(k))$ is called Second-Level Prediction Model.

As far as Double-Level Prediction Model goes, The base steps of the multi-Level Recursive prediction Method, as reference (3) mentioned, go as follow.

1) According to observing data, we use proper distinguishing method to estimate the unkown parameter $\theta(k)$ of the model (3).

for exampple, the following Recursive Algorithm can be applied.

$$\hat{\theta}(k)=\hat{\theta}(k-1)+\frac{1}{1+\varphi(k)^t\varphi(k)}\varphi(k)\{Y(k)-\varphi(k)^t\hat{\theta}(k-1)\}$$

Where $\hat{\theta}(k)$ expresses the estimating value of $\theta(k)$

2) According to the estimated value sequence $\hat{\theta}(0)$, $\hat{\theta}(1),\cdots\hat{\theta}(N)$, we can set up the Second-Level Prediction Model $\hat{\theta}(k+1)=M(\hat{\theta}(k))$, Where N, is the present moment, and

$\hat{\theta}(k)=\{\hat{\theta}(k),\hat{\theta}(k-1),\cdots\hat{\theta}(k-q)\}$

3) Through model $M(\hat{\theta}(k))$, the prediction value $\hat{\theta}(N+1)$, $\hat{\theta}(N+2)$, $\cdots\hat{\theta}(N+h)$ of the time varying parameter $\theta(k)$ can be obtained. Where h is prediction step length.

4) According to input value series $\hat{U}(N+1-p)$, $\hat{U}(N+2-p),\cdots\hat{U}(N+h-p)$, which were given before and using the First-level Prediction Model $\hat{Y}(N+h)=\varphi(N+h)^t\hat{\theta}(N+h)$, the prediction value $\hat{Y}(N+1),\hat{Y}(N+2),\cdots\hat{Y}(N+h)$, can be obtained. Where

$\varphi(N+1)=\{-Y(N),\cdots -Y(N+1-h),\hat{U}(N+1-p),U(N-p),\cdots U(N-p+1-m)\}$

$\varphi(N+2)=\{-\hat{Y}(N+1),-Y(N),\cdots-Y(N+1-p-h),\hat{U}(N+2-p),\hat{U}(N+1-p),\cdots U(N+2-m-p)\}$

II. Application Examples

1) The prediction results of Heilong jiang province gross output value of industry

The following numbers are known historical data, which were computed according to fixed price. The unit is RMB 100 million yuan.

year	gross output value of industry	investment in fixed assets
1971	131.9	5.94
1972	138.4	6.67
1973	149.2	8.46
1974	159.9	8.96
1975	177.7	10.97
1976	187.9	8.26
1977	204.3	8.89
1978	222.1	13.25
1979	236.7	12.70
1980	249.1	15.90
1981	256.5	14.05
1982	274.1	18.68
1983	295.6	24.39
1984	325.0	26.10
1985	360.8	28.65
1986	389.4	34.09
1987	432.4	41.22
1988	482.1	43.57
1989	511.3	38.97
1990	519.7	45.39

According to the formulas which were given above, the recent years prediction values are as follows.

year	true value	prediction value	error rate(%)
1987	432.4	423.7	−2.0
1988	482.1	461.6	−4.3
1989	511.3	505.0	−1.2
1990	519.7	555.6	+6.9

2) About the prediction results of Harbin city gross output value of industry

The following numbers are known historical data.

The price is also fixed price. The unit is RMB 100 million yuan.

year	gross output value of industry	year	gross output value of industry
1971	29.2	1981	56.7
1972	30.8	1982	60.1
1973	33.2	1983	65.9
1974	33.3	1984	75.7
1975	37.4	1985	88.6
1976	37.5	1986	96.1
1977	42.0	1987	108.2
1978	47.5	1988	125.6
1979	52.2	1989	127.7
1980	55.7	1990	125.5

According to the formulas which were given above, the recent years prediction values are as follows

year	true value	prediction value	error rate(%)
1987	108.2	107.1	−1.0
1988	125.6	116.0	−7.6
1989	127.7	125.9	−1.4
1990	125.5	134.5	+7.2

The prediction results above-mentioned express that either in Heilongjiang province or in Harbin city, the error rates in 1987 and 1989 were within 1 percent to 2 percent. But the error rates in 1988 and 1990 were between 4 percent to 7 percent. The industry increase in 1988 and 1990 were in unusual situation. In 1988, economy was inflation. Total amount was uncontroled. Industry increased over usual years. The output value of industry than that of last year were 11.5 percent and 16 percent respectively. So 4.3 percent and 7.6 percent errors occured. In 1990, demand total amount was controled. The unreasonable economy structure was adjusted. The increase speed of industry decreased. So around 7 percent errors occured. These facts show that using Double-level Prediction Model and Multi-Level Recursive Prediction Method to predict the dynamic of regional industry economy system is suitable in normal years. But the thorough-going and concrete analysis should be done in unusual years.

Reference

(1) Han Zhigang, Multi-Level Recursive Identification Method. Chinese Journal of Automation. Vol1, No. 1 (1989)

(2) Han Zhigang, The Multi-Level Recursive Method and Application, Science press. Beijing. 1989.

(3) Han Zhigang, On New Method of Dynamic System Prediction. ACTA Automatica Sinica. Vol 9, No. 3 (1983).

Computing the steady state distribution of networks with negative and positive customers

J.M. Fourneau
L.R.I., bât 490
Université de Paris Sud, 91405 Orsay, France

Abstract

Recently a new class of queueing networks with "negative and positive" customers was introduced and shown to have product form solution. Positive customers are identical to the usual customers of a queueing network, while a negative customer deletes a customer when it arrives to a non empty queue. As the customer flow equation is a non linear fixed point equation, the stability problem remains open. In this paper we present an algorithm to compute the steady state distribution of these networks and check the stationnarity of the queues. We also establish stability of the system as the proof of the algorithm implies the existence of a solution.

I Introduction

In a recent paper [2], a new class of queueing networks in which customers are either "negative" or "positive" was introduced and was shown to have a product form solution. Positive customers enter a queue and receive service as ordinary queueing network customers, while a negative customer will delete a positive customer.

Models with positive and negative customers can be used to represent different systems. The first study on network with negative customers was motivated by the analogy with neural networks [3], [4]. Queues with positive and negative customers may model systems of objects exchanging positive and negative signals. Another possible application is to represent multiple resource systems. Positive customers can be considered to be resource requests, while negative customers can correspond to decisions to cancel these requests.

Unlike BCMP or Jackson networks [1], the customer flow equations of this model are non-linear. Therefore it is not easy to determine the necessary and sufficient conditions for the existence of a solution inside the domain of stability of the network (i.e. the conditions under which the stationary solution exists [3]). Furthermore even if we know the form of the solution, we do not know how to compute it except in the case of feed-forward networks.

In this paper we propose an algorithm to compute the steady state distribution of this new type of networks. In section II we introduce the product form theorem shown in [2]. In section III we present an algorithm to iteratively compute q_i the load of service center i and check the stationnarity of the Markov process. We prove the algorithm, evaluate the complexity of one iteration. The proof of the algorithm implies that there always exists a solution to the fixed point system we consider. Our results may be useful either in Neural Networks theory or in Queueing Network theory.

II The Model

We consider networks with an arbitrary number N of queues. All service centers have exponential service time distributions. In this model, when a negative customer arrives to a non-empty queue it deletes a positive customer. A negative customer arriving to an empty queue just vanishes. Negative customers do not receive service.

External arrival streams to the network are independent Poisson processes. We denote by Λ_i the external arrival rate of positive customers to queue i and by λ_i be the external arrival rate of negative customers to queue i. The services are exponential with mean μ_i.

Each customer may change service centers and nature at the completion of its service. These movements of customers between queues are represented by a Markov chain. Negative customers leaving a queue (either after having destroyed a positive customer, or after finding an empty queue) disappear from the network.

A positive customer which leaves queue i (after finishing service) goes to queue j as a positive customer with probability $P^+[i,j]$, or as a negative customer with probability $P^-[i,j]$. It may also depart from the network with probability $d[i]$. We shall denote the state of the queueing network by the vector $x = (x_1, \ldots, x_N)$ where x_i represents the state of service center i and is as usual the number of customers in queue i. We denote by $|x_i|$ the total number of customers in queue i. Let $\Pi(x)$ denote the stationary probability distribution of the state of the network, if it exists. The following result shows the product form solution of the networks being considered (see [2] for a proof).

Theorem 1 *For an open network of stations the stationary steady state probabilities are given by* $\Pi(x) = \prod_{i=1}^{N} g_i(x_i)$; *where each* g_i *is a function depending on the service center, if the system of non linear equations (1), (2), (3) :*

$$q_i = \frac{\Lambda_i + \Lambda_i^+}{\mu_i + \lambda_i + \lambda_i^-} \tag{1}$$

$$\Lambda_i^+ = \sum_{j=1}^{N} P^+[j,i]\mu_j q_j \tag{2}$$

$$\lambda_i^- = \sum_{j=1}^{N} P^-[j,i]\mu_j q_j \tag{3}$$

have a solution such that $\forall i \quad 0 < q_i < 1$ λ_i^- *and* Λ_i^+ *are the arrival rates of negative and positive customers arriving to station* i *from other stations in the network. Furthermore the* $g_i(x_i)$ *in (2) have the following form :* $g_i(x_i) = (1 - q_i)[q_i]^{|x_i|}$

Obviously, the customer flow equations (Eq (1), (2), (3)) for this type of networks are non linear. So the existence of their solutions is a difficult problem except for feed-forward networks (i.e. the networks are directed acyclic graphs). In this case it is possible to compute the load of the queues in the topological order of the graph. The network stability conditions and the existence of the solutions to customer flow equations introduce some mathematical problems which had been partially solved in [3] using an homotopy function [6]. In the next section, we state an algorithm to compute the load $(q_i)_{i=1..N}$. Therefore we prove that there always exists a solution to this fixed point system of equations.

III The Algorithm

Instead of solving the initial problem introduced in theorem 1, we design an algorithm to solve the following fixed point system :

$$\begin{aligned} \Lambda_i^+ &= \textstyle\sum_{j=1}^{N} P^+[j,i]\mu_j q_j \\ \lambda_i^- &= \textstyle\sum_{j=1}^{N} P^-[j,i]\mu_j q_j \\ q_i &= Min\left(1, \frac{\Lambda_i + \Lambda_i^+}{\mu_i + \lambda_i + \lambda_i^-}\right) \end{aligned} \tag{4}$$

Note that the stationnarity condition is now included into the fixed point system since q_i is positive and smaller than or equal to 1.

If for any i q_i is equal to 1, then we show that queue i is not stable. Otherwise, the solution of the second problem is obviously the solution of the first problem. Furthermore uniqueness of the solution follows from the Markovian framework of the model.

Description

We consider, for any queue index i, six sequences of real numbers $[\overline{q_i}]_k$, $[q_i]_k$, $[\overline{\lambda_i^+}]_k$, $[\lambda_i^+]_k$, $[\overline{\lambda_i^-}]_k$, $[\lambda_i^-]_k$ defined by induction (on k) as follows

$$\begin{aligned}
[\overline{\lambda_i^+}]_k &= \Lambda_i + \sum_{j=1}^{N} P^+[j,i]\mu_j[\overline{q_j}]_k \\
[\lambda_i^+]_k &= \Lambda_i + \sum_{j=1}^{N} P^+[j,i]\mu_j[q_j]_k \\
[\overline{\lambda_i^-}]_k &= \lambda_i + \sum_{j=1}^{N} P^-[j,i]\mu_j[\overline{q_j}]_k \\
[\lambda_i^-]_k &= \lambda_i + \sum_{j=1}^{N} P^-[j,i]\mu_j[q_j]_k \\
[\overline{q_i}]_{k+1} &= min(1, [\overline{\lambda_i^+}]_k/(\mu_i + [\overline{\lambda_i^-}]_k)) \\
[q_i]_{k+1} &= min(1, [\lambda_i^+]_k/(\mu_i + [\lambda_i^-]_k))
\end{aligned} \tag{5}$$

with the following initial values $[q_i]_0 = 0$ and $[\overline{q_i}]_0 = 1$, and we iterate the computations of the sequences. The following theorem states that under certain conditions these sequences lead to the solution of the problem.

Theorem 2 *If for any service center i, one of the following assumptions is satisfied*

- *the probability that a positive customer leaves the station to go outside is strictly positive (i.e. $d[i] > 0$).*
- *there is a strictly positive probability that a customer, either positive or negative, joins a queue j where the rate of negative customers coming from the outside is strictly positive. (i.e. $P^+[i,j] + P^-[i,j] > 0$ and $\lambda_j > 0$)*

then the algorithm converges to N values of the load $(q_i)_{i=1..N}$ which is a solution of this new fixed point system

Proof of the algorithm

Lemma 1 *For all i the sequences satisfy the following relations*

$$[q_i]_k \le [\overline{q_i}]_k \quad \text{and} \quad [\lambda_i^+]_k \le [\overline{\lambda_i^+}]_k \quad \text{and} \quad [\lambda_i^-]_k \le [\overline{\lambda_i^-}]_k \tag{6}$$

The sequences $[\overline{q_i}]_k$, $[\overline{\lambda_i^+}]_k$ and $[\overline{\lambda_i^-}]_k$ are non increasing and the sequences $[q_i]_k$, $[\lambda_i^+]_k$ and $[\lambda_i^-]_k$ are non decreasing.

Proof: by induction on k.

Lemma 2 *The sequence defined by the differences between the two sequences $[\overline{q_i}]_k$ and $[q_i]_k$ is positive non increasing. Furthermore under the same assumptions as theorem 2, the summation over the queues of these differences multiplied by the rate of service of the queue μ_i is decreasing to zero.*

Proof: We have to consider two cases according to the value of the lower sequence $[q_i]$.

- If for some I $[q_I]_l$ is equal to 1 then obviously $[\overline{q_I}]_l$ is also equal to 1. And for all $k > l$ the two sequences stay equal to 1. The solution of the fixed point system is on the boundary of the set of definition for the load q_i. The station is not stable.
- On the other hand, if $[q_i]_k$ is strictly smaller than 1 we easily give a bound on the difference between $[\overline{q_i}]_k$ and $[q_i]_k$. Let $[\Delta\lambda_i^+]_k$ $[\Delta\lambda_i^-]_k$ and $[\Delta q_i]_k$ be the new sequences defined by the difference between the upper and lower sequences for iteration k,

$$\begin{aligned}
[\Delta\lambda_i^+]_k &= [\overline{\lambda_i^+}]_k - [\lambda_i^+]_k \\
[\Delta\lambda_i^-]_k &= [\overline{\lambda_i^-}]_k - [\lambda_i^-]_k \\
[\Delta q_i]_k &= [\overline{q_i}]_k - [q_i]_k
\end{aligned} \tag{7}$$

Using that $[q_i]_k$ is smaller than 1 for all k, we get .

$$[\Delta q_i]_{k+1} \le [\overline{\lambda_i^+}]_k/(\mu_i + [\overline{\lambda_i^-}]_k) - [\lambda_i^+]_k/(\mu_i + [\overline{\lambda_i^-}]_k) \tag{8}$$

And after some algebraic manipulations we obtain :

$$\sum_{i=1}^{N} \mu_i[\Delta q_i]_{k+1} \le \sum_{j=1}^{N} \mu_j[\Delta q_j]_k \sum_{i=1}^{N} (P^+[j,i] + P^-[j,i]) \frac{\mu_i}{\mu_i + \lambda_i} \tag{9}$$

Let $\epsilon_j = \sum_{i=1}^{N} (P^+[j,i] + P^-[j,i]) \frac{\mu_i}{\mu_i + \lambda_i}$, and let ϵ be the maximum over the queue index j of the values ϵ_j. Then,

$$\sum_{i=1}^{N} \mu_i[\Delta q_i]_{k+1} \le \sum_{j=1}^{N} \mu_j[\Delta q_j]_k \epsilon_j \le \epsilon \sum_{j=1}^{N} \mu_j[\Delta q_j]_k \tag{10}$$

Therefore the sequence of real values $\sum_{i=1}^{N} \mu_i[\Delta q_i]_k$ is bounded by a geometric sequence of rate ϵ. Note now that the assumptions of theorem 2 may be restated as $\epsilon < 1$, this implies that the sequence $\sum_{i=1}^{N} \mu_i[\Delta q_i]_k$ is decreasing to zero.

Clearly lemma 2 proves that the algorithm leads to the solution of the fixed point system. If the solution is such that all the components q_i are smaller than 1 0, then it is also a valid solution for the initial problem : q_i is the load of station i in the network.

Otherwise if a component q_i is on the boundary of the domain then this implies that the lower bound $[\lambda_i^+]_k/(\mu_i + [\lambda_i^-]_k)$ is greater than 1. So there is no solution of the fixed point system inside the domain and queue i is unstable.

Thus the proof of theorem 2 is complete and the solutions of the fixed point system (8) give the loads q_i of the queues in the network or show that some queues are unstable.

Each iteration of the algorithm consists in the computation of the new values of the sequences and the new value of the difference to check the accuracy. There are $6N$ new values to compute and four of them (the most expensive in term of computations) require a linear number of operations. Therefore the complexity of each iteration is quadratic.

IV Conclusion

In this paper we prove the stability (the existence of a solution) of the fixed point system of equations (1) (2) and (3) using an algorithm to obtain the solution. The uniqness of this solution in the domain of ergodicity is obtained because of the Markovian framework. The solution is accurate as we obtain lower and upper bound for the load of the queues. Furthermore the algorithm is efficient as the difference between upper and lower bounds is decreasing quicker than a geometric sequence.

Since we derived this algorithm, the model of positive and negative customers was improved to take into account customer class. It is shown in [5] that networks with multiple classes of positive customers and negative customers also have product form solution and the fixed point system is slightly different. Application of the algorithm to this new fixed point system is straightforward.

References

[1] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios *Open, closed and mixed networks of queues with different classes of customers*, ACM Journal, Vol. 22, No 2, p248-260, April 1975

[2] E. Gelenbe *Product form networks with negative and positive customers*, To appear in Journal of Applied Probability, 1989

[3] E. Gelenbe *Stability of Random Neural networks*, Neural Computation, 1990.

[4] E. Gelenbe *Random Neural Networks with Negative and Positive Signals and Product Form Solution*, Neural Computation, 1990

[5] J.M. Fourneau, E. Gelenbe, R. Suros *Multiclass G-Networks*, EHEI Report, Submitted for publication, 1990.

[6] C.D. Garcia, W.I. Zangwill, *Pathways to solutions, fixed points, and equilibria*, Prentice Hall, Englewood Cliffs, N J., 1981.

COMPUTATIONAL METHODS IN THE OPTIMIZATION OF A STATIONARY (s,S) INVENTORY PROBLEM

B. D. Sivazlian
Department of Industrial and Systems Engineering
The University of Florida
Gainesville, Florida 32611

ABSTRACT - We consider a periodic review, single product, single installation inventory control system with probabilistic demand operating in the steady state under the (s,S) policy. A unified methodology is provided to generate five computational techniques for the determination of the optimum decision variables s and Q=S-s so as to minimize the steady-state total expected cost of operation per period. The computational methods discussed are: an analytic method, a numerical method, an analogue method, an asymptotic method and an approximate method.

1. THE MATHEMATICAL FORMULATION

Consider a periodic review inventory control system with immediate delivery of orders where unfilled demands are backlogged. The demand during successive periods is described by a sequence of independently and identically distributed continuous random variables with probability density function $\phi(\xi)$ and distribution function $\Phi(\xi)$ $(0<\xi<\infty)$, both assumed to be Laplace transformable.

We shall assume that the procurement cost has two components: a fixed component K (K > 0) and a variable component c.z, where z is the total quantity ordered. In addition, we shall assume that a cost L(y) is incurred during each period where y is the initial stock level at the beginning of a period following a decision. The ordering or stock replenishment policy will be of the (s,S) type, that is, if at the end of a period the stock level is x<s, a quantity S-x is ordered, otherwise no order is placed.

It can be verified [1] that if Q = S-s the expression for the total expected cost per period takes the form

$$g(S,Q) = \frac{K + L(S) + \int_0^Q L(S-\xi)\psi(\xi)d\xi}{1 + \int_0^Q \psi(\xi)d\xi} + cD \qquad (1)$$

where $$\psi(u) = \phi(u) + \int_0^u \phi(u-v)\psi(v)dv \qquad (2)$$

and that the minimum value of this cost at optimality is:

$$g(S^*,Q^*) = L(s^*) + cD$$

where $$D = \int_0^\infty \xi\phi(\xi)d\xi$$

Let h denote the unit holding cost and p the unit penalty cost. Since y can take on negative values, the expression for L(y) becomes:

$$L(y) = \begin{cases} h\int_0^y (y-\xi)\phi(\xi)d\xi + p\int_y^\infty (\xi-y)\phi(\xi)d\xi & y \geq 0 \quad (3) \\ p\int_0^\infty (\xi-y)\phi(\xi)d\xi & y < 0 \quad (4) \end{cases}$$

Theorem

The optimal values of Q and s, Q* and s*, satisfy the following set of equations

$$M(s^*,Q^*) = K \quad \text{and} \quad \left.\frac{\partial M(s^*,x)}{\partial x}\right|_{x=Q^*} = 0 \qquad (5)$$

where $$\bar{M}(s^*,z) = \mathcal{L}(M(s^*,x)) = \frac{\mathcal{L}(L(s^*) - L(s^*+x))}{1 - \mathcal{L}(\phi(x))} \qquad (6)$$

2. COMPUTATIONAL METHODS

2.1 An Analytic Method

In this method, the function $M(s^*,x)$ is first determined explicitly by inverting analytically expression (6); the system of equations (5) is then solved for s* and Q*. Although this method is quite general, nevertheless, it is convenient only when the expressions for L(y) and $\phi(x)$ are rather simple In addition and quite often, the inversion of $M(s^*,z)$ can be extremely tedious. Using for L(y) the expression given in (3) with $\phi(x)$ taken as the gamma density function with parameters μ and r:

$$\phi(x) = \frac{(\mu x)^{r-1}}{\Gamma(r)}\mu e^{-\mu x}, \quad r > 0, \ \mu > 0 \qquad (7)$$

we obtain for r = 1

$$M_1(s^*,x) = -\frac{h}{2\mu}(\mu^2x^2 + 2\mu x) + \frac{e^{-\mu s^*}(h+p)}{2\mu}2\mu x \qquad (8)$$

2.2 A Numerical Method

This procedure consists in inverting $\mathcal{L}(M(s^*,x))$ in (6) numerically for a range of values of s*. The value of s* and x are then selected to satisfy the system of equations (5). To illustrate the approach, we consider the case of linear holding and shortage cost when the distribution of demand is gamma as given by (7).

To reduce the total number of input variables, we perform a dimensional analysis. We first note that the problem involves five input variables (p, h, K, μ and r) and two output variables which are the optimum decision variables s* and Q*. We can show that it is possible to express the two dimensional output factors μs^* and μQ^* as a function of the three dimensional

input factors $\frac{P}{h}$, $\frac{K\mu}{h}$ and r. In the sequel we shall express the dimensional factors by

$$A = \frac{P}{h}\ ,\ B = \frac{2K\mu}{h}\ ,\ r\ ,\ X = \mu s^*\ \text{and}\ Y = \mu Q^* \qquad (9)$$

Equation (5) can be written as:

$$\int_0^Y m_0(X,t)\ dt = \frac{B}{2} \qquad (10)$$

$$m_0(X,0) = 0 \qquad (11)$$

where $m_0(X,t)$ is the derivative of the function $M(s^*,x)$ defined in (6), appropriately scaled. Figure 1 illustrates a typical output.

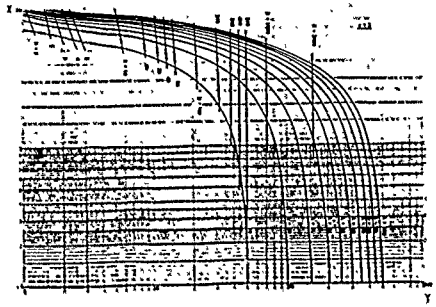


Figure 1. Optimal values of $X = \mu s$ and $Y = \mu Q$ when demand is gamma distributed of order r = 5 (A = p/h, B = 2Kμ/h)

2.3 An Analogue Method

This method is based on the analogy that equation (6) bears with a feedback control system. It is possible to design an electrical analogue circuit to simulate the above system. By setting s as a parameter, conditions (5) can be made to be satisfied.

Consider the demand per period to be of the gamma type given by (7) with h = 1, p = 10, μ = .4, and r = 4. The plot of the output function $m(s,x) = M'(s,x)$ for various values of s is presented in Figure 2.

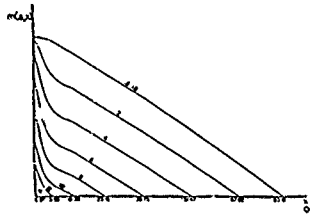


Figure 2. Hybrid computer results for m(s,x) when r = 4 and μ = .4.

2.4 An Asymptotic Method

Assume L(y) is given by (3). Denote $H(x) = 1-\Phi(x)$, then

$$\mathcal{L}\{M(s^*,x)\} = -\frac{h}{z^2[1-\bar{\phi}(z)]} + (h+p)\frac{\mathcal{L}\{H(s^*+x)\}}{z[1-\bar{\phi}(z)]}$$

and

$$M(s^*,x) = -\mathcal{L}^{-1}\left(\frac{h}{z^2[1-\bar{\phi}(z)]}\right) + (h+p)\int_0^x \mathcal{L}^{-1}\left(\frac{\mathcal{L}\{H(s^*+x)\}}{1-\bar{\phi}(z)}\right)du \qquad (12)$$

For large x, we can write:

$$\bar{\phi}(z) = 1 - zD + \frac{1}{2}z^2(D^2+\sigma^2) + o(z^2)$$

where $D < \infty$ and $\sigma^2 < \infty$ denote the finite mean and variance of the R.V.ξ. So that for large x, the first term of (12) becomes:

$$\mathcal{L}^{-1}\left(\frac{h}{z^2[1-\bar{\phi}(z)]}\right) = \frac{h}{2D}x^2 + \frac{h}{2}\left(1+\frac{\sigma^2}{D^2}\right)x + o(1)$$

A similar expression can be obtained for the second term. We obtain

$$Q^* \doteq \sqrt{\frac{2KD}{h}} \text{ and } \int_{s^*}^{\infty} H(v)dv \approx \frac{D}{h+p}\left[\frac{h}{2}\left(1+\frac{\sigma^2}{D^2}\right)+Q^*\right]$$

2.5 An Approximate Method

If we assume that $M(s^*,x)$ is expandable as a power series of x, then for small x we have:

$$M(s^*,x) \doteq a_0 + a_1 x + a_2 x^2$$

a_0, a_1 and a_2 can be evaluated by making use of the known inherent characteristics of the function $M(s^*,x)$. The following identity is obtained, where $R(\cdot)$ is some remainder term:

$$\left[\frac{2K}{Q^*} + L'(s^*)\right]x - \left(\frac{K}{Q^{*2}} - \frac{1}{2}[L''(s^*)+L'(s^*)\ \phi(0)]\right)x^2 + [R_1(x)-R_2(x)] = 0$$

This identity yields the following approximate system of equations in s^* and Q^* for small Q^*

$$L'(s^*) = -\frac{2K}{Q^*} \quad \text{and} \quad L''(s^*) = \frac{2K}{Q^{*2}} + \frac{2K}{Q^*}\phi(0)$$

This method in conjunction with the asymptotic method could be used to provide bounds on s^* and Q^*.

3. REFERENCES

[1] Scarf, H. E., D. M. Gilford and M. W. Shelley (eds), Multistage Inventory Models and Techniques, Chapter 1, Stanford University Press, Stanford, CA, 1963.

OPTIMAL CONTROL WITH A CONTINUOUS TIME MODEL OF THE ITALIAN ECONOMY[1]

Fusari Angelo
Istituto di Studi per la Programmazione Economica (ISPE)
Corso Vittorio Emanuele,282, 00186 Rome, Italy.

Abstract-The paper applies the maximum principle to a model of the Italian economy specified as a stock-flow continuous time macrodynamic model and estimated by a full information maximum likelihood (FIML) estimator.

I. INTRODUCTION

The purpose of this study is to outline at an aggregate level some optimizing strategies and policy design aimed at coping with the spontaneous tendencies in Italy of income distribution and public finance seriously troubled by a large and incessantly growing public deficit. The research hinges upon a model principally devoted to make evident the relations connecting the public and private sectors of the Italian economy (1).

II. SPECIFICATION OF THE OBJECTIVE FUNCTION

Three kinds of objective function (a, b and c) will be employed:

The first (a) maximizes the level of private sector value added and of gross investment in manufacturing industry and services and minimizes the fluctuations of the above investments from a trend, using as instruments the interest rate and currency. In symbols, this objective function takes the following form:

$$\max \int_{81}^{86} \{ q_y \left[\log PROD_{sp(t)} + \log O_{1(t)}\right] + q_I \log I_{1(t)} - q_{It}\left[\log I_{1(t)} - \log \bar{I}_{1(t)}\right]^2 - k_{\dot{m}}\left[\dot{m}(t) - \bar{\dot{m}}(t)\right]^2 - k_i \left[i(t) - \bar{i}(t)\right]^2 \} dt$$

PROD indicates the private sector labour proctivity; O_1 is employment in industry and services; I_1 stays for gross investment in manufacturing industry and services in real terms; $\dot{m}$ indicates the variation rate of currency; i is domestic nominal interest rate for long term government bonds. The dash on the variable indicates a desired (or ideal) value. All desired values are represented by trends. To be precise:

$\log \bar{I}_{1(t)} = \log I_{1(t_*)} + 0.01t$; $\bar{\dot{m}}(t) = 0.03 + 0.01t$;

$\bar{i}(t) = 0.0191 - 0.001t$; (in quarterly growth rates)

We have employed the following weights:

$q_y = 1.0$; $q_I = 0.5$; $q_{I(t)} = 0.5$

A second form of objective function (b) has been obtained by dropping in the objective function specified above the term which minimizes the fluctuations of gross investment about its trend and adding, as control variable, a term which minimizes the deviations of money wage rate in the private sector from a desired path of this variable; therefore, this new objective function also includes the expression:

$$-k_w \{\log w_{psp(t)} - \log \bar{w}_{psp(t)}\}^2$$

Where $\log \bar{w}_{psp(t)} = \log w_{psp(t_*)} + 0.03t$

Finally, the third kind of objective function (c) adds in the previous one a new target which minimizes the deviations of the public debt from a desired path of this variable. Besides, it substitutes the wage policy by two new control variables which minimize the deviation of public sector expenditures and of public sector entrances from some hypothetical ratios between these variables and private sector nominal value added. So, the new objective function replaces the wage policy, in the previous one, by the following expressions:

$$-k_u \{\log U_{T(t)} - \log PROD_{sp(t)} - \log O_{1(t)} - \log P_{(t)} - \delta\}^2$$

where $\delta = 0.6$ is a desired ratio (expressed in logarithms) between public expenditure (U_T) and private sector nominal value added, and $PROD_{sp}$, O_1, P are the variables determining private nominal value added.

$$-k_E \{\log E_{T(t)} - \log PROD_{sp(t)} - \log O_{1(t)} - \log P_{(t)} - \tau\}^2$$

Where $\tau = 0.75$ is a desired ratio (expressed in logarithms) between public entrances and private nominal value added, and E_T indicates the public sector revenue.

Besides, we add the new target:

$$-q_B \{\log B_{(t)} - \log \bar{B}_{(t)}\}^2$$

Where: $\log \bar{B}_{(t)} = \log B_{(t_*)} + 0.05t$ and $q_B = 0.5$. B is the total amount of the public debt.

We refer the explorations to a planning horizon going from 1981 to 1986.

III. BRIEF DISCUSSION OF SOME RESULTS.

The optimization results , being expressed in logarithms, allow us to take directly the variation rates of each variable. It may be useful to expound a table containing the average rates of growth of some more relevant variables in the planning horizon, obtained using each one of the three kinds of objective function specified above.

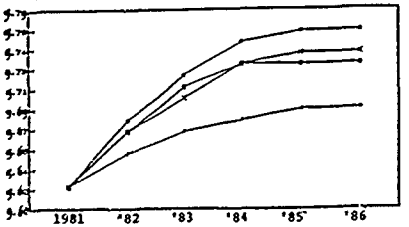
[1] The computation programs and the continuous methodology used in this application are due to Dr. C.R. Wymer.

Table 1 Percent yearly average growth rates of the objectives.

Objective variables	Object. func.(a)	Object. func.(b)	Object. func.(c)	Forec. value
O_1	2.41	2.71	2.07	1.36
$PROD_{sp}$	4.96	4.57	4.40	3.84
I_1	9.55	8.30	8.99	5.85
B			8.60	15.55

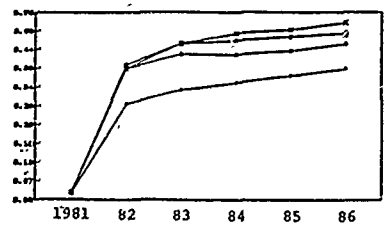
Table 1 shows that the adoption of each one of the objective functions leads to a remarkable improvement of the targets, with respect to the neutral solution. When the objective function (b) is adopted, employment growth appears nearly doubled in confront of its forecast value probably owing to the use, in this case, of wage rate as control variable. But even in the other cases employment growth is remarkable. By adding the growth rates of employment and productivity, we find that the yearly average rate of growth of the value added in the private sector is almost 2 percent more than its forecast value. Also the dynamics of private investment shows a marked push. Finally, the objective function (c) including public deficit among its targets, shows: a) a considerable reduction in the dynamics of this variable and, b) that the use of public entrance and expense as policies does not yield a great benefit to employment and productivity growth. It may be useful to expound the trajectories of some variables.

Employment in industry and services (O_1)

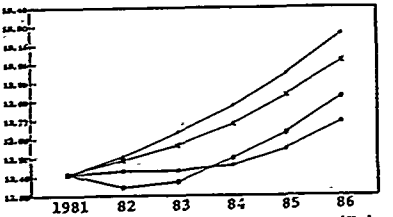


• Forecast value × Optimal value (a)
▫ Optimal value (b) ◦ Optimal value (c)

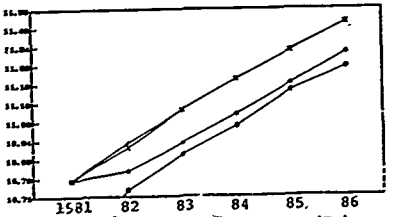
Gross real invest. in manuf. and services (I_1)



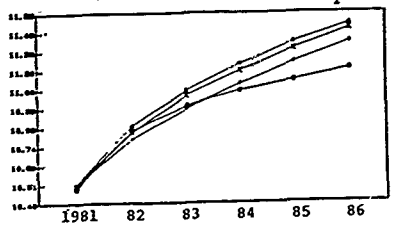
Total amount of public debt (B)



Public sector expenditure (U_1)



Public sector revenue (E_1)



It is interesting to observe the trajectories of public debt and of the policy instruments. We can see that the fulfilment of an incisive wage policy, as the objective function (b) does, determines a sharp reduction of public debt growth. For their part, public finance policies postulated by the objective function (c) cause a drastic fall of public debt at the beginning of the period, which is followed later by a considerable growth of this variable, the trend of which remains much lower than the neutral solution. This result is due to a strong reduction of public expenditure, while fiscal policy shows a moderate trend. For space limitations, we avoid further comments on the above graphs and conclusions.

REFERENCES

(1) Fusari A. (1990), "A macrodynamic model centered on the interrelationships between the public and private sectors of the Italian economy", New York Economic Review.

(2) Wymer, C.R. "Trans, Resmul, Contnest, apredic computer programs and relative manuals".

A STRATEGY FOR SOCIOECONOMIC SIMULATION OF EUROPE 1992

Joan Raul Walter
California State University, Dominguez Hills
Computer Information Systems, SBS D-325
Carson, CA 90747 U.S.A.

Abstract - Socio-economic Simulation [SES] is a multiple decision-maker process allowing competitive or collaborative behavior among persons sharing a common simulated economic environment. Discussed here are the problems of handling judgmental information in the modeling of consumption and investment. This is implemented with the application of Computer Assisted Software Engineering [CASE] diagrams and the use of linked spreadsheets.

I THE PROBLEM OF JUDGMENTAL INFORMATION

A problem often encountered in the modeling of economic events is the inability to include judgmental information together with the quantitative data. Human reaction to changes in the simulated environment can be of greater importance than the numerical precision of the estimates afforded by the model's equations.

One way to surmount the problem of judgmental information is to actually embed human decision-makers as role players within a simulation exercise or "game." Thus a simulation model contains not only data, a set of equations and procedures, but furthermore a network of participants. In this network, participants may be subject to partial ordering to reflect hierarchies of organizations. Although their essential role may be competitive, the decision-makers are empowered to help each other understand and solve common problems for their mutual benefit through their negotiated sharing of information normally kept confidential. Consequently, the simulation is defined by who the participants are, in addition to what they do.

Participants' actions are of special significance with this approach. Their behavioral patterns are recorded by the computer system and become an integral part of the modeling process. In subsequent exercises the system's adaptation to decision-makers' behavioral patterns is incorporated within the simulated economic environment, for example, by using the empirical data of participants' responses to change the bias of one or more random number generators in the model.

The result is a truly interactive process, where the model's properties influence participants' behavior. This behavior, in turn, causes changes in the policies and procedures comprising the model. Various examples of the application of this strategy toward modeling the 1992 economic unification of Europe have been tested.

II MODELING CONSUMPTION AND INVESTMENT

In transnational organizations, decision makers need to decide not only whether to build or buy technology, but also where that is to be done. At the national level, consumption and investment can be described:

$C = C(0)+mpc*(Y-T(0)+t*Y-TP(0)+p*Y)$, and

$I = I(0)-c*r$, with the variables:

I(0)	Autonomous investment
c	Marginal efficiency of capital
T(0)	Autonomous tax revenues
t	Marginal tax rate
TP(0)	Autonomous transfer payments
p	Relationship between transfer payments and income
e	Exchange rate
C(0)	Autonomous consumption
mpc	Marginal propensity to consume
Y	Country A's national income
Y(B)	Country B's national income

The constructs presented below are new approaches to business simulation. The simulation serves as a "Gedankenexperiment", to detect incipient difficulties before any actual decisions are taken.

III APPLICATION OF CASE DIAGRAMS

Fig. 1 below is a diagram of flow of funds, patterned after a Data Flow Diagram [DFD]. This diagram shows part of the relationship among a transnational firm, a transnational bank and a national treasury.

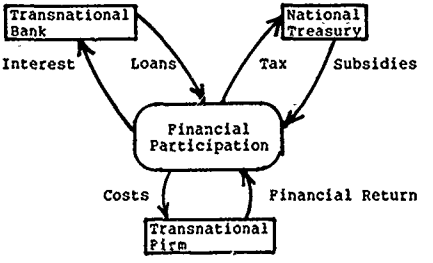


Fig. 1. Flow of Funds

Fig. 2 below shows a decomposition diagram, protraying the hierarchical relationship among the firm, the bank and the treasury, below the director of the simulation.

Fig. 2. Hierarchy of simulation participants.

IV USE OF LINKED SPREADSHEETS

This approach uses hidden spreadsheets linked with a master spreadsheet. The master spreadsheet is accessible to all participants. Its primary contents are numbers which are the results of calculations. It may also contain formulas, but only those formulas intended to be public knowledge (e.g., VAT = .15 indicating a value-added tax rate). To obtain the results, the master spreadsheet draws upon hidden spreadsheets in the custody of participating decision-makers. Formulas in the hidden spreadsheets are privy only to the simulation director and the specific decision-maker to whom that spreadsheet has been allocated by the director.

Fig. 3 below corresponds to the decomposition diagram of Fig. 2.

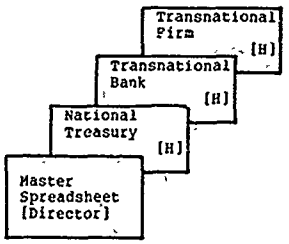


Fig. 3. Linking of 3 Hidden Spreadsheets

V REFERENCES

1. [Anonymous], Supercalc5 User's Guide, Computer Associates, San Jose, CA, 1988.
2. C.S. Greenblat and R.D. Duke, Gaming - Simulation: Rationale, Design and Applications, John Wiley and Sons, New York, 1975.
3. C. Lopilato, Europe 1992: A Computer Spreadsheet Simulation Game, California State University, Dominguez Hills, Carson, CA, 1989.
4. C. Lopilato and J.P. Walter, "An Artificial Intelligence Approach to the Simulation of Europe 1992", in AI and Simulation: Theory and Applications, edited by Wade Webster and Ranjeet Uttamsingh, Society for Computer Simulation, San Diego, CA, 1990.
5. W. A. McEachern, Economics: A Contemporary Introduction, South-Western, Cincinnati, OH, 1988.
6. S.L. Nelson, "A Spreadsheet Business Planner", Lotus, April 1990, pp. 37-38.
7. P. Rosenkranz, [1979]. An Introduction to Corporate Modeling, Duke University Press, Durham, NC, 1979.
8. A.N. Schreiber, [1970]. Corporate Simulation Models, University of Washington, Seattle, 1970.
9. J.P. Walter, PRECOS - Un Nouveau Systeme Ordinateur pour la Simulation Socio-economique, Faculty of Sciences, University of Paris, 1971.
10. D. Warsh, The Idea of Economic Complexity, Viking Press, New York, 1980.

DEALING WITH UNCERTAINTY IN MULTIPLE OBJECTIVE DECISION SUPPORT

C. Henggeler Antunes (1,3) and João N. Clímaco (2,3)
(1) Department of Electrical Engineering ; (2) Faculty of Economics - University of Coimbra - 3000 Coimbra - Portugal
(3) INESC, Rua Antero de Quental, 231 cave - 3000 Coimbra - Portugal

Abstract - This paper presents an interactive approach to deal with the inherent uncertainty and imprecision arising in problems where multiple conflicting objectives exist. The uncertainty and imprecision are associated not just with the initial data and the imperfection of the model as a representation of reality, but also with the decision maker's preference structure which evolves throughout the interactive decision process as more knowledge about the problem is gathered. An interactive approach aimed at evaluating the stability of compromise solutions is proposed, which is based on the representation of the solutions in the weight space.

I. INTRODUCTION

Most complex real-world problems are characterized by multiple, conflicting and incommensurate objectives. Mathematical models as well as the perception of the problems by decision makers become more realistic if several objectives are considered explicitly. In a multiple objective context, the concept of optimal solution gives place to the concept of nondominated solution (feasible solution for which there are no other feasible solution that improves an objective function value without worsening at least other objective function value). These decision problems entail tradeoffs among the objectives, in order to get a satisfactory compromise solution from the set of nondominated solutions.

Different methods for tackling multiple objective problems exist, using different solution techniques and requiring distinct degrees of involvement of the decision maker (DM). In interactive methods phases of decision involving the DM are alternated with phases of computation. These methods reflect a diversity of strategies for carrying out the search for nondominated solutions and differ in the type of information required from and presented to the DM. The DM intervenes in the solution search process by inputing information into the procedure which in turn is used to guide the search process (thus minimizing the computational effort) in order to compute a new solution which more closely correspond to his preferences. For a review see [4].

The interactive process plays a significant role to the enhancement of the knowledge acquisition process, by improving skills to gain new insights into the problems. These may then be used to express new preferences and progressively narrowing down the scope of the search, thus minimizing the computational burden. The support of a learning process is an essential step to construct more realistic representations of the DM's preferences, aimed at capturing the lack of surety and/or preference changes that may arise as the interactive decision process proceeds. The possibility of performing some kind of stability or sensitivity analysis is thus a very important component of interactive decision aid computational tools. On the one hand, it enables the DM to consider modifications of the original data, concerning uncertainty, inaccuracy and changes associated with the input data, as well as the imprecisions stemming from the modelling phase. On the other hand, it enables the DM to exploit changes of his preference structure, thus accommodating his lack of surety in specifying new search directions required in a next computation phase of the interactive process.

The approach proposed to evaluate these effects on nondominated solutions is integrated in the operational framework of the TRIMAP package. The TRIMAP method [1] is at the core of an interactive package aimed at supporting the DM in the progressive and selective learning of the set of nondominated solutions in multiple objective linear programming (MOLP) problems. The capabilities of the computer environment (namely graphical displays) are exploited in order to provide the DM an user-friendly man-machine interface.

II. AN INTERACTIVE DECISION SUPPORT TOOL FOR MOLP

A. Multiple Objective Linear Programming

Multiple objective linear programming is concerned with problems of the type:

$$\text{"Max"}\ f(x) = Cx \qquad (1)$$
$$\text{s.t.}\quad x \in F$$
$$F = \{x \in R^n : Ax = b,\ x \geq 0\}$$

where $C = [c_1, c_2, \ldots, c_p]^T$ (objective function matrix), c_k, $k = 1,2,\ldots,p$, are n-dimensional vectors, A (technological matrix) is a mxn matrix and b (RHS) is a m-dimensional vector. "Max" denotes the operation of finding the set of nondominated solutions (in a maximizing sense).

The set of efficient (Pareto optimal, noninferior) solutions is defined by $F_E = \{x \in F : f(x^1) \geq f(x) \Rightarrow x^1 \in F\}$

where $f(x^1) \geq f(x)$ iff $f(x^1) \geq f(x)$ and $f(x^1) \neq f(x)$, and $f(x^1) \geq f(x)$ iff $f_k(x^1) \geq f_k(x)$, $k = 1,2,\ldots,p$.

In this case we would say that $f(x)$ is nondominated when $x \in F_E$. In general, whereas the concept of nondominance refers to the objective function space, the concept of efficiency refers to the decision variable space.

A process of computing efficient solutions consists in solving scalar problems consisting of a weighted sum of the objective functions as in (2). The admissible set of weights is defined by $\Lambda = \{\lambda : \lambda \in R^p, \Sigma \lambda_k = 1, \lambda_k > 0, k = 1,2,\ldots,p\}$. By fixing a set of weights $\lambda \in \Lambda$ a weighted linear function to be maximized over F is obtained

$$\max \sum_{k=1}^{p} \lambda_k f_k(x) \qquad (2)$$
$$\text{s.t.}\quad x \in F, \lambda \in \Lambda$$

It has been proved that $x \in F$ is an efficient basic solution to (1) if and only if it is an optimal solution to (2). The graphical display of the set Λ which leads to each efficient solution can be achieved by means of the decomposition of the weight space Λ. The weight space may be perceived as the space of the relative importance of the objective functions.

From the simplex tableau corresponding to an efficient basic solution to (2) the corresponding Λ set is given by $\Lambda^T W \geq 0$. $W = C_B B^{-1} A - C$ is the reduced cost matrix (one line for each objective function, where the element w_{kj} is the marginal rate of change of objective function $f_k(x)$ caused by the introduction of one unit of variable x_j into the basis), B is the basis and C_B is the submatrix of C corresponding to the basic variables.

The region comprising the set of weights corresponding to a nondominated extreme solution (region where $(\lambda^T W \geq 0, \lambda \in \Lambda)$ is consistent) is called indifference region [3]. The DM can thus be indifferent to all the combinations of weights within it, because they lead to the same nondominated solution. The analysis of the indifference regions in the weight space is a valuable tool in "learning the shape" of the nondominated solution space. The boundaries between contiguous indifference regions represent the nonbasic efficient variables (those which when introduced into the basis lead to an adjacent efficient extreme point through an efficient edge). A common boundary between two indifference regions means that the corresponding efficient solutions are adjacent and connected by an efficient edge. If a point $\lambda \in \Lambda$ belongs to several indifference regions this means that they correspond to efficient solutions lying on the same face.

B. The TRIMAP Package

TRIMAP [1] is an interactive MOLP package which is aimed at aiding the DM in the progressive and selective learning of the set of nondominated solutions. TRIMAP combines three main basic procedures: weight space decomposition, introduction of constraints on the objective function space, and introduction of constraints on the weight space. Moreover, TRIMAP enables the introduction of additional limitations on the objective function values to be translated into the weight space. The dialogue with the DM is made mainly in terms of the objective function values (which is the type of questions that do not require an excessive effort from the DM, unlike assessing marginal rates of substitution or other forms of preference elicitation). In general, the weight space is used in TRIMAP as a valuable means for collecting and presenting the information.

The main purpose of TRIMAP is to provide a progressive and selective filling of the weight space. In each dialogue phase the DM must indicate whether the study of solutions corresponding to not yet searched regions is of interest. This enables the DM to perform a "strategic search" of the feasible polyhedron and prevents the exhaustive search in regions with close objective function values (which often arise in real-world problems). The interactive process only ends when the DM considers to have gathered "sufficient knowledge" about the set of nondominated solutions which enables him to make a decision. The stopping rule is thus the "satisfaction" of the DM rather than the checking for convergence of any utility function.

TRIMAP is dedicated to problems with three objective functions (p-3). Although this is a limitation, it allows for the use of graphical means which are particularly suited for the "dialogue" with the DM. In order to enhance his capabilities in processing the information, the DM is offered a flexible and user-friendly environment, always keeping the control over the solution search process. The computer interface is mainly based on: a menu bar at the top of the screen lists the titles of the available pulldown menus, grouping the actions available to the user, thus not occupying screen space and not requiring command memorization; overlapping windows are used by the program for displaying tabular and graphical information to the user, pictorial controls provide the user an intuitive way for specifying his preferences; dialogue boxes are used whenever some further information is needed before a given command can be carried out, being also useful for conveying status information to the user. The introduction of constraints on the objective values and their translation into the weight space enables the dialogue with the DM to be done in terms of the objective values accumulating the resulting information in the weight space.

In each interaction of TRIMAP two graphs are presented to the DM. The first one is the weight space displaying the regions corresponding to each nondominated extreme solution already known. Eventual constraints on the variation of the weights are also presented, whether they are directly introduced into the weight space or result from additional constraints imposed on the objective function values. The second graph displays the nondominated solutions already computed, projected on a plane of the objective function space. Further details about the working of the TRIMAP method as well as the main features of its computer implementation can be found in [1].

III. EVALUATING THE IMPACT OF CHANGES IN DATA AND DM'S PREFERENCES

In single objective linear programming, sensitivity (stability, post-optimal) analysis deals with computing ranges on the variation of some initial data such that the optimal basis remains optimal for the perturbed problem. The concept of optimal solution (in general unique) gives place in multiple objective programming to the concept of efficient solution (in general many, even if only extreme points are considered). Moreover, changes in the underlying DM's preference structure as a result of the information gathered throughout the interactive process must be taken into account. The definition of stability analysis in a multiple objective context is not uniformly adressed in the literature (see [2]).

In this paper linear parametrizations of the objective function matrix and the resource availability (right-hand side of the constraints) will be considered. An analysis based on the weight space is proposed in the framework of the TRIMAP package, which enables to present graphical information to the DM in a way which promotes rapid comprehension.

A. Changes in the objective function matrix

A perturbation of the objective function matrix depending on a scalar parameter is considered. $C(y) = C + y D$, where $D = [d_1 d_2 d_3]^T$ is a constant matrix and y is a scalar parameter ($y \in R$).

Let us use the partitions $A = [B, N]$, $C = [C_B, C_N]$, $D = [D_B, D_N]$, $x = [x_B, x_N]$, where B, C_B and D_B correspond to the basic variables x_B, and N, C_N and D_N correspond to the nonbasic variables x_N. Let J_N be the set of nonbasic variables.

For a given efficient solution, the parameter y does not have influence on the primal feasibility ($B^{-1} b \geq 0$). For the solution to remain efficient the dual feasibility must be satisfied (with $\lambda \in \Lambda$):

$$\lambda^T [C_B(y) B^{-1} N - C_N(y)] \geq 0$$

$$\Leftrightarrow \lambda^T (C_B B^{-1} N - C_N) + y \lambda^T (D_B B^{-1} N - D_N) \geq 0$$

We then have

$$\lambda^T (C_B B^{-1} N_j - C_{N_j}) + y \lambda^T (D_B B^{-1} N_j - D_{N_j}) \geq 0 \quad , j \in J_N ,$$

and $\lambda \in \Lambda$.

(where N_j, C_{N_j} and D_{N_j} denote the j^{th} column of matrix N, C_N and D_N, respectively).

Let us consider $\alpha_j = C_B B^{-1} N_j - C_{N_j}$ and $\beta_j = D_B B^{-1} N_j - D_{N_j}$ and write the dual feasibility condition as $\lambda^T \alpha_j + y \lambda^T \beta_j \geq 0$, $j \in J_N$, and $\lambda \in \Lambda$.

The basis B will correspond to an efficient extreme point if and only if the following system is consistent.

$$\begin{cases} \sum_{k=1}^{3} \lambda_k (\alpha_{kj} + y \beta_{kj}) \geq 0 & j \in J_N \\ \lambda \in \Lambda \end{cases}$$

The problem of stability analysis of a given efficient solution will consist in determining the range on parameter y for which the solution is "still preferred" according to a "pattern of preferences" represented by the indifference region of this solution for the unperturbed (y=0) problem. The problem consists then in finding the minimum (y_{min}) and maximum (y_{max}) values of the scalar parameter y, for each efficient solution selected by the DM, so that the intersection between the unperturbed and the perturbed indifference region is nonempty. This means finding y_{min} and y_{max} subject to

$$\sum_{k=1}^{3} \alpha_{kj} \lambda_k \geq 0 \ , \ j \in J_N \qquad (3a)$$

$$\sum_{k=1}^{3} \alpha_{kj} \lambda_k + y \sum_{k=1}^{3} \beta_{kj} \lambda_k \geq 0 \ , \ j \in J_N \qquad (3b)$$

$\lambda \in \Lambda$.

For three-objective problems, it is possible to use the graphical displays of the indifference regions an an aid to the DM for evaluating the stability of selected solutions to the uncertainty, imprecision and changes associated with the data and his preference structure. Let $S_C^+ = \{j \in J_N :$ $\sum_{k=1}^{3} \beta_{kj} \lambda_k > 0$, in (3a)$\}$ and $S_C^- = \{j \in J_N : \sum_{k=1}^{3} \beta_{kj} \lambda_k < 0$, in (3a)$\}$.

From (3b) it follows: $y \geq -\sum_{k=1}^{3} \alpha_{kj} \lambda_k / \sum_{k=1}^{3} \beta_{kj} \lambda_k$, for $j \in S_C^+$,

and $y \leq -\sum_{k=1}^{3} \alpha_{kj} \lambda_k / \sum_{k=1}^{3} \beta_{kj} \lambda_k$, for $j \in S_C^-$

So, considering each $j \in \hat{N}$ separately the problem would reduce to a linear fractional programming problem to be optimized subject to (3a). As it is well known, this problem has its optimum at an extreme point of the feasible region. The simultaneous consideration of constraints (3b) (which are nonlinear in y and λ_k) makes the problem more complex.

The extreme points of each (unperturbed) indifference region are known explicitly (given by (3a), as a by-product of the decomposition of the weight space in TRIMAP). Let E_s be the set of extreme points of the indifference region (a convex polygon) corresponding to the (basic) efficient solution s: $E_s = \{(\lambda_1^r, \lambda_2^r, \lambda_3^r), r = 1, \ldots, R_s, \sum_k \lambda_k^r = 1 \ , \ \lambda_k \geq 0 \ , \ k = 1,2,3\}$

The dual feasibility condition (3b) can then be tested exhaustively at $\lambda \in E_s$ for computing the range of variation of the parameter y. The range on the parameter y which satisfies (3) for a given nondominated solution s can then be computed in the following manner

(let $y_j = -\sum_{k=1}^{3} \alpha_{kj} \lambda_k / \sum_{k=1}^{3} \beta_{kj} \lambda_k$):

$$y_{min} = \min \{ y_j : j \in S_C^+, \sum_{k=1}^{3} \alpha_{kj'} \lambda_k + y_j \sum_{k=1}^{3} \beta_{kj'} \lambda_k \geq 0 ,$$

$\forall j' \in J_N, j' \neq j, \forall \lambda \in E_s\}$

$y_{min} = -\infty$ if $S_C^+ = \emptyset$

$$y_{max} = \max \{ y_j : j \in S_C^-, \sum_{k=1}^{3} \alpha_{kj'} \lambda_k + y_j \sum_{k=1}^{3} \beta_{kj'} \lambda_k \geq 0 ,$$

$\forall j' \in J_N, j' \neq j, \forall \lambda \in E_s\}$

$y_{max} = +\infty$ if $S_C^- = \emptyset$

The computations involved are very simple and do not require solving any auxiliary programming problem.

The indifference regions of the perturbed problem change continuously (in size and form) with changes of the parameter y. Note that due to the nonlinearity in y and λ_k, y_{min} and y_{max} which satisfy (3) may not correspond to an extreme point of the unperturbed indifference region. These cases are easily detected by the algorithm to construct the indifference regions from the system of inequalities resulting from dual feasibility. In these circumstances a bissection technique is used to compute the actual values y_{min} or y_{max} (or both) for which the intersection reduces to one point. This implies reconstructing the perturbed indifference region for each value of y generated by the bissection technique and calculating the region where it intersects the unperturbed indifference region (intersection of two convex polygons). From a computational point of view, these are easy tasks in the framework of TRIMAP (for instance, of a much lower level of complexity than solving linear programming problems with parameters in the technological matrix).

The area occupied by each indifference region in the weight space may be understood as a measure of the robustness of the corresponding nondominated solution to the variation of the weights.

The DM can also visualize the changing of the indifference regions in an interactive manner by varying y dynamically. This type of visual interactive stability analysis can similarly be performed for a group of nondominated solutions, namely those defining nondominated edges or faces (which are easily identified in the weight space), by previously calculating their ranges on the parameter y and considering their intersection. The possibility of changing dynamically the value of y also enables to have a visual interactive information about regions of the weight space where "holes" appear, meaning that new efficient extreme solutions to the perturbed problem can be searched in those regions.

B. Changes in the resource availability

Let us consider a perturbation of the resource availability depending on a scalar parameter:

$\underline{b}(t) = \underline{b} + t \ \underline{h}$, where $\underline{h}$ is a constant m-dimensional vector and t is a scalar parameter ($t \in R$).

For a given efficient basis (corresponding to an efficient extreme point), the parameter t does not have influence on the dual feasibility. For the solution to remain efficient the primal feasibility must be satisfied.

$$B^{-1} \underline{b}(t) \geq \underline{0}$$

$$\Leftrightarrow \quad B^{-1} \underline{b} + t B^{-1} \underline{h} \geq \underline{0}$$

For variations of the RHS the feasible polyhedron changes itself. So the indifference regions do not change in a "smooth" manner (as in the case of objective function matrix perturbation) with the variation of the parameter t, but they change "suddenly" as efficient extreme points appear or disappear because of the changes in the RHS

Let $S_b^+ = \{i : B_{i.}^{-1} \underline{h} > 0\}$ and $S_b^- = \{i : B_{i.}^{-1} \underline{h} < 0\}$

and $t_i = -B_{i.}^{-1} \underline{b} / B_{i.}^{-1} \underline{h}$, ($B_{i.}^{-1}$ denotes the i^{th} row of B^{-1}).

t_{max} and t_{min} can be determined in the following way:

$$t_{min} = \max \{ t_i , i \in S_b^+ \}, \ t_{min} = -\infty \text{ if } S_b^+ = \emptyset$$

$$t_{max} = \min \{ t_i , i \in S_b^- \}, \ t_{max} = +\infty \text{ if } S_b^- = \emptyset$$

For t in the range defined in this way, the perturbed and unperturbed indifference regions are the same.

IV. AN ILLUSTRATIVE EXAMPLE

Let us consider the illustrative MOLP problem:

$$\text{"Max"} \begin{bmatrix} 3 & 1 & 2 & 1 \\ 1 & -1 & 2 & 4 \\ -1 & 5 & 1 & 2 \end{bmatrix} \underline{x} \text{ s.t.} \begin{bmatrix} 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 1 & 2 & 3 & 4 \end{bmatrix} \underline{x} \leq \begin{bmatrix} 60 \\ 60 \\ 50 \end{bmatrix}, \underline{x} \geq \underline{0},$$

$\underline{x} \in R^4$. Let $D = \begin{bmatrix} 1 & -1 & 2 & 0 \\ 4 & -1 & 0 & -1 \\ 0 & 1 & 3 & 2 \end{bmatrix}$ be a constant matrix.

TRIMAP first computes the solutions that optimise each objective function: solutions 1 and 3 optimise f_1 and f_3, respectively; solutions 2 and 4 are alternative optima with respect to f_2. In fig. 1b, the black square represents the ideal solution (the one that would optimise all the objective functions simultaneously), and the number after the identification of the solution is the value of f_3. Suppose that the DM then decided to compute some more solutions in order to gather some knowledge about the set of nondominated solutions. The situation where all nondominated extreme solutions are known is displayed in fig. 1 (in general, this is not the aim of TRIMAP). The analysis of the two graphs shows that 3 nondominated faces exist, defined by solutions (1,7,8), (2,4,6) and (2,6,5,7,8). The points lying on the face defined by solutions (3,5,6) are dominated by the points on the edges (3,5) and (5,6)

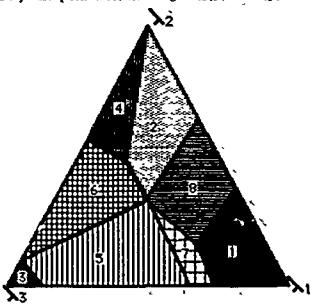


Fig. 1a - The weight space graph

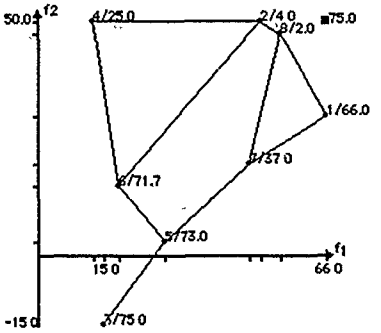


Fig. 1b - The projection of the objective space on f_1f_2

Let us suppose that the DM wants to investigate the range of variation of the parameter y with respect to solution 8. The weight spaces corresponding to $y_{min} = 0.393$ and $y_{max} = 0.750$, for which the intersection between the unperturbed (see fig. 1) and the perturbed indifference regions is nonempty (reducing to one point) are displayed in fig. 2. In fig. 2a all the solutions are still nondominated, and the face (3,5,6) also becomes nondominated. In fig. 2b, solutions 2, 3 and 4 are no longer nondominated, and only the face (1,7,8) and the edge (5,7) are still nondominated.

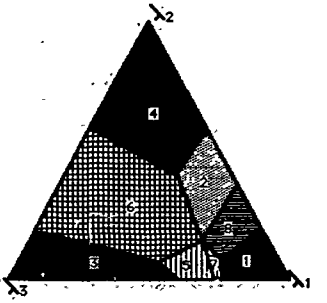


Fig. 2a - The weight space corresponding to y_{min}

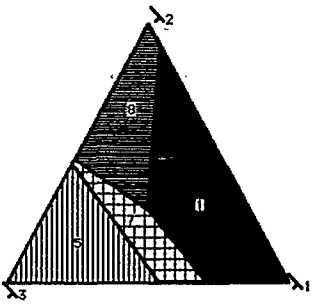


Fig. 2b - The weight space corresponding to y_{max}

V. CONCLUSIONS

An interactive approach to deal with the uncertainty and imprecision arising in MOLP problems has been presented. The uncertainty and imprecision are associated both with the initial data and the DM's preference structure which evolves throughout the interactive decision process.

The weight space is used as a graphical means for collecting and presenting the information to the DM. The stability of nondominated solutions to changes in the objective function matrix and in the resource availability have been analyzed in terms of the intersection of the unperturbed and perturbed indifference regions (perceived as a "pattern of preferences" of the DM).

REFERENCES

[1] J. Climaco and C. H. Antunes, "Implementation of a user friendly software package - a guided tour of TRIMAP". Mathematical and Computer Modelling 12, 1299-1309 (1989).

[2] T. Gal and K. Wolf, "Stability in vector maximization - a survey" European Journal of Operational Research 25, 169-182 (1986).

[3] J. Kornbluth, "Duality, indifference and sensitivity analysis in multiple objective linear programming". Operational Research Quarterly 25, 599-614 (1974).

[4] D. Vanderpooten and Ph. Vincke, "Description and analysis of some representative interactive multicriteria procedures" Mathematical and Computer Modelling 12, 1221-1238 (1989).

PETRI NETS: COMBINATORIAL APPROACH AND MODELLING FLEXIBLE MANUFACTURING SYSTEMS

ALEXANDER A.OREKHOV
Informatics and Computing Chair of the
Ukrainian Academy of Sciences
Prospekt Akademika Glushkova, 20
Kiev-207, 252207, USSR

Abstract- The new approach to investigation of Petri nets based models of FMS is presented in this paper. Its main idea is to describe a given PN as a discrete dynamic system and to evaluate properties of this system. This leads to determination of a new relations between structural properties of PN and ability to solve effectively some sheduling problems for FMS. It is also showed that incidence matrix of PN gives important information about behaviour of this net and hence about properties of system being investigated.

1.INTRODUCTION

We shall use such denotations: PN is a quadruple

$$N=(m,n;A^+,A^-)$$

where m is a number of places enumerated from 1 to m and designated by $p_1, \ldots, p_m$, n is a number of transitions enumerated from 1 to n and designated by $t_1, \ldots, t_n$, matrices A^+ and A^- (both have size $m \times n$ and their elements are nonnegative integers) describe relations between places and transitions:

$$a^+_{ij} \ (a^-_{ij})$$

is a weight of arc connecting t_j and p_i (p_i and t_j, respectively). Initial marking is denoted by m_0, process of PN firing is described by the chain

$$m_0 \to t_{j(1)} \to m_{(1)} \to t_{j(2)} \to m_{(2)} \to \ldots .$$

Succession of firing transitions

$$t_{j(1)} t_{j(2)} \ldots$$

gives a word s of free prefix language $P^f(m_0)$ of PN N and Parikh images $t=PI(s)$ of all elements of $P^f(m_0)$ form a set of reachable states $T(m_0)$. It should be pointed out that every reachable marking m corresponds to s as

$$m=(A^+-A^-)PI(s)+m_0,$$

where $m_0 \to s \to m$. Let $t^0, t^1, \ldots, t^k$ be a set of integer nonnegative n-vectors connected by the recurrent system of unequalities

$$\begin{cases} t^l \geq t^{l-1} \\ A^+t^{l-1} \geq A^-t^l - m_0 \end{cases} \qquad (1)$$

where $l=1,\ldots,k$ and $t^0=0_n$.

2.RELATIONS BETWEEN WORDS OF PETRI NET AND SOLUTIONS OF SYSTEM (1)

Lemma. 1. For every solution of system (1) there exists a word $s=s_1s_2\ldots s_k$ of $P^f(m_0)$ such that for every $l=1,\ldots,k$

$$t^l=PI(s_1s_2\ldots s_l). \qquad (2)$$

2. For every finite s belonging to $P^f(m_0)$ there exists a number k and a set

$$(t^0, t^1, \ldots, t^k) -$$

solution of system (1) such that for every $l=1,\ldots,k$ s can be broken into such subwords $s_1, s_2, \ldots, s_l$ that (2) holds.

3. For every positive integer k there exists an isomorphism between subset

$P^f(m_0)$ including all words of length k and set of those solutions of system (1) which are related by the equation

$$\|t^l\|_E - \|t^{l-1}\|_E = 1 \text{ for every } l=1,\ldots,k. \ \square$$

This lemma gives ground for investigation of properties of PN N by means of analysis of system (1) which can be treated as a simulation tool for real system.

Let us discuss two combinatorial problems concerning Petri net theory. The first one is formulated as follows.

Suppose that PN N describes some FMS and such a task is formulated: find

$$\max (c,t^k) \text{ if } (t^0, t^1, \ldots, t^k) \text{ belongs to a set of solutions of system (1).} \qquad (3)$$

Solution of this task will give optimal in definite sense shedules of functioning of FMS given.

Theorem 1([1]). 1. Task (1), (3) is strongly NP-complete.

2. If PN N is a finite automaton or a marked graph then task (1),(3) can be solved by time which is evaluated as a polynomial function of m, n and k. $\square$

Proof of part 1 of this theorem is based on the reduction of a task of finding a hamiltonian path in arbitrary oriented graph to task (1),(3) having special structure. Proof of part 2 includes nontrivial transformations of system (1) to show that matrix of limitations of this system is totally unimodular.

This theorem implies that for sheduling FMS we have to use heuristics in the case of general PN and strict algorithms developed in theory of combinatorial optimisation if PN is an ordinary one.

The second problem being investigated concerns relations between structural properties of its incidence matrix $A=A^+-A^-$. It should be pointed out that there exists theoretically and practically effective algorithm for search of PN invariants ([2]); this means that such a problem is of real interest when investigating and designing FMS.

We shall analyse system $Ax \geq 0_m$ where x is nonnegative real n-vector. We shall say that condition (4) holds if

set of nonnegative solutions of system of unequations $Ax \geq 0_m$ coincides with set of nonnegative solutions of system of equations $A'x=0_m$ where A' is received from A by deleting those rows of A which correspond to excessive limitations of system $Ax \geq 0_m$ where all x are nonnegative (4)

Theorem 2([3]). If condition (4) holds then set of reachable states $T(m_0)$ of PN is semilinear and free prefix language $P^f(m_0)$ of this net is regular for every initial marking m_0.□

Corollary. Fair (in the sence of definition of paper [4]) PNs have semilinear sets of reachable states $T(m_0)$ for every m_0. □

Note that the validity of condition (4) for orbitrary PN can be recognized with the help of algorithm proposed in ([2]).

In conclusion it should be pointed out that search of PN invariants is realised as a program for IBM/PC which showed good results for PNs with 50 or less transitions (if T-invariants were being sought). Now it is being modified to fit tasks with 1000 transitions.

References

1. A.A.Orekhov. Matem. Metody Obrab. Inform. i Upravl.Moscow, M.Ph.-T.I Publ., 1988. P. 111-116 (in russian).
2. A.A.Orekhov. Kibernetika, 1988. N. 3. P. 102-103 (in russian).
3. A.A.Orekhov. (unpublished) 1989.
4. T.Murata, Z.Wu. J.Franklin Inst., 1985. V. 320. N. 2. P. 63-82.

FINITE ELEMENT SOLUTION OF THE REVERSE SPAGHETTI PROBLEM

Madana M. G. Gopal
EASI
Auburn Hills, Michigan 48057 U.S.A.

Xavier J. R. Avula
University of Missouri-Rolla
Rolla, Missouri 65401 U.S.A.

Abstract: In this study, a mathematical model of a flexible paper strip moving axially in a fluid medium under the inertial and steady, incompressible aerodynamic effects is developed. The strip is treated as a linear elastica which is capable of undergoing large displacements. The model equations are derived on the basis of Newton's laws of motion and moment equilibrium. These equations are solved using a semi-discrete finite element analysis to produce the strip's trajectory and configuration in spatio-temporal domain. In the absence of the aerodynamic forces these equations degenerate to the case of the so-called "reverse spaghetti" problem of an elastica emerging from a horizontal guide. Further, the motion of the strip is experimentally investigated, and its profile and the leading edge trajectory are compared. The results are in good agreement. The method of analysis and the results are applicable to the study of motion of thin paper strips such as documents in a document-processing machine.

I. INTRODUCTION

Industrial applications of axially moving materials such as traveling strings, band saws and magnetic tapes are numerous [1]. In most cases, these materials are continuous and move on multiple supports. However, there are situations such as paper transport in copying and check processing machines in which the material has finite length and is not supported on both its leading and trailing edges. Most often it is supported on one edge and the leading edge is free.

Carrier [2] determined the motion of a dangling, inextensible string that was drawn upward through a hole in a rigid wall. He called this type of problem "the spaghetti problem." The "reverse spaghetti problem" which deals with the drooping motion of an axially-moving elastica issuing from a horizontal guide was solved by Mansfield and Simmonds [3], where only the inertial effect on the motion of the material was considered. This analysis is applicable only when the elastica has a high weight-to-stiffness ratio and low velocities.

The present study differs from the above problems in that the moving material has very low bending rigidity and is propelled at high velocities, as in the case of a high trajectory of the leading edge, which is one of the main concerns in a high-speed environment. The present effort addresses this issue with a view of assisting the designers of high-speed equipment that handles axially moving materials.

The axially moving material is assumed to be a thin, highly flexible strip. The governing equations of motion for the strip capable of undergoing large deformations are derived using Newton's laws of motion and moment equilibrium. The effect of aerodynamic forces is brought in through an "added mass," a concept developed from the theory of hydrodynamics. Application of these equations leads to a system of coupled nonlinear partial differential equations. A semi-discrete finite element approximation is used to solve these equations subjected to appropriate initial and boundary conditions.

II. MATHEMATICAL MODEL

The geometrical representation of an emerging elastica is shown in Figure 1. An element of a thin flexible strip with various forces acting on it are shown in Figure 2.

Equation of motion in the x-direction is

$$\frac{\partial \bar{H}}{\partial X} - f_d(X,T) + f_{SU}\delta(X-L)\cos\beta(L,T) - f_{IM}\cos\beta = mb\frac{\partial^2 \bar{X}}{\partial T^2} \quad (1)$$

where $\delta(\bullet)$ is a dirac-delta function.

The equation of motion in the y-direction is

$$\frac{\partial V}{\partial X} - mbg + z(X,T) + f_{SU}\delta(X-L)\sin\beta(L,T) - f_{IM}\sin\beta = mb\frac{\partial^2 \bar{Y}}{\partial T^2} \quad (2)$$

where $f_d(X,T)$, $z(X,T)$, f_{SU}, and f_{IM} are the drag, lift, leading edge suction, and initial impulse, respectively.

Using the notation

$$\frac{\partial(\cdot)}{\partial T} = (\cdot)_T\ ,\quad \frac{\partial^2}{\partial T^2}(\cdot) = (\cdot)_{TT} \text{ etc.}$$

$$\bar{V}(X,T) = mbg\int_L^X d\xi - \int_L^X z(\xi,L)\,d\xi + \int_L^X f_{SU}\delta(\xi,L)\sin\beta(L,T)\,d\xi + \int_L^X f_{IM}\sin\beta(\xi,T)\,d\xi + mb\int_L^X \bar{Y}_{TT}(\xi,T)\,d\xi \quad (3)$$

where

$$\int_L^X z(\xi,T)\,d\xi = -m\mu b\left[\int_L^X \bar{Y}_{TT}(\xi,T)\,d\xi + 2V\int_L^X \bar{Y}_{\xi T}(\xi,T)\,d\xi + V^2\int_L^X \bar{Y}_{\xi\xi}(\xi,T)\,d\xi\right] = -m\mu b\{Q_{TT} + 2V(Q_{XT}(X,T) - Q_{XT}(L,T)) + V^2(\sin\beta(X,T) - \sin\beta(L,T))\} \quad (4)$$

$$\int_L^X f_{SU}\delta(\xi,L)\sin\beta(L,T)\,d\xi = \frac{mb\mu}{2}\{Q_{XT}^2(L,T) + 2VQ_{XT}(L,T)\sin\beta(L,T) + V^2\sin^2\beta(L,T)\}\sin\beta(L,T) \quad (5)$$

$$\int_L^X f_{IM}\sin\beta(\xi,T)\,d\xi = mb\mu V^2 k_9 T^{-1/2}\int_L^X \sin\beta(\xi,T)\,d\xi = mb\mu V^2 k_9 T^{-1/2}\{Q_X(X,T) - Q_X(L,T)\} \quad (6)$$

$$mb\int_L^X \bar{Y}_{TT}(\xi,T)\,d\xi = mbQ_{TT}(X,T) \quad (7)$$

Introducing the above integrals, the equations of motion become

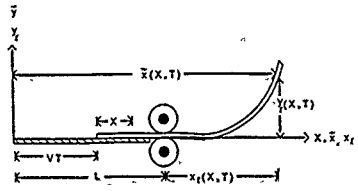
$$\bar{H}(X,T) = mb\left[D_1 - \frac{\mu}{2}(Q_{XT}^2(L,T) + 2VQ_{XT}(L,T)\sin\beta(L,T) + V^2\sin^2\beta(L,T))\cos\beta(L,T) + k_9\mu V^2 T^{-1/2}(P_X(X,T) - P_X(L,T)) + P_{TT}\right] \quad (8)$$

where

$$D_1 = \mu\int_L^X (k_4 + k_5 T^{3/2}(\xi,T))\,d\xi \quad (9)$$

and

$$\bar{V}(X,T) = \frac{1}{2}\left[q(X-L) + (\lambda+\mu)Q_{XT} + 2\mu V\,(Q_{XT}(X,T) - Q_{XT}(L,T)) + \mu V^2(\sin\beta(X,T) - \sin\beta(L,T))\,\frac{\bar{x}}{2}\,(Q^2_{XT}(L,T) + 2VQ_{XT}(L,T)\sin\beta(L,T) + V^2\sin^2\beta(L,T))\sin\beta(L,T) + \chi_0\,\mu V^{2-1/2}(Q_X(X,T) - Q_X(L,T))\right] \quad (10)$$



INSIDE THE GUIDE. $0 < X < L\text{-}VT$ OUTSIDE THE GUIDE. $L\text{-}VT < X < L$

$\bar{x}(X,T) = X + VT$ $\qquad$ $\bar{x}(X,T) = L + x_f(X,T)$

$\bar{y}(X,T) = 0$ $\qquad$ $\bar{y}(X,T) = Y_f(X,T)$

Figure 1. Geometrical representation of an emerging elastica.

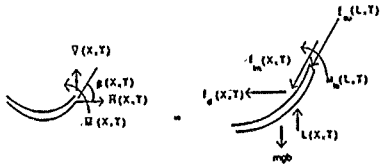


Figure 2. Internal and external forces and moments acting on an emerged portion of a flexible strip.

The boundary conditions are

$$\text{at } X = L - VT\,,\ \beta = Q_X = 0\,,\ P_X = L$$
$$\text{at } X = L\,,\ P = Q = 0\,,\ \text{and}\ \beta_X = \frac{M_{le}}{B^e} \quad (11)$$

With the added mass ratio $\mu = 0$ the above equations degenerate to the "non-aerodynamic" case and yields the so-called "reverse spaghetti problem."

III. THE FINITE ELEMENT SOLUTION

Using the finite difference for the time domain and finite elements for space discretization, a finite element grid is constructed on [0,1] with β, p, and q as the nodal degrees of freedom.

$$\text{Let } P = L^2p \text{ and } Q = L^2q \quad x = L(1-s) \quad (12)$$

$$\text{and } T = (L/V)t \quad (13)$$

$$\text{where } s = 1 - (X/L) \text{ or } (ds/dx) = -(1/L)$$

An approximate solution is constructed on an element of length "h" as

$$\beta^h(\sigma) = \sum_{j=1}^{N} \beta_j^n(\tau)\phi_{1,j}(\sigma) \quad (14)$$

$$p^h(\sigma) = \sum_{j=1}^{N} p_j^n(\tau)\phi_{2,j}(\sigma) \quad (15)$$

and

$$q^h(\sigma) = \sum_{j=1}^{N} q_j(\tau)\phi_{3,j}(\sigma) \quad (16)$$

where $\phi_{i,j}$ are piecewise linear or quadratic polynomials which satisfy the essential boundary conditions.

$$\phi_{1,i}(1) = \phi_{2,i}(0) = \phi_{3,i}(0) = 0 \quad (17)$$

Using backward difference for time, the weak formulation of the problem is performed. Omitting the details due to length limitation, the stiffness matrix expressed in local coordinates after assembly leads to a banded, unsymmetric global stiffness matrix. For a linear element, the element stiffness matrix is 6x6 and for quadratic element it is 9x9. The method of solution discussed here pertains to modified Newton's method. For convenience, the time domain is approximated by backward difference and it is conditionally stable. Thus for accuracy, the time step $\Delta\tau$ should be smaller than "h," the grid spacing used in finite element discretization. Choosing a smaller time step also ensures that no oscillations in the solution are introduced. In most of the cases considered, i.e. for problems without aerodynamic forces, the time step $\Delta\tau = 0.02$ or $\Delta\tau = 0.01$ was found sufficient to produce good results.

IV. REVERSE SPAGHETTI PROBLEM: DISCUSSION

The present day document processing equipment depending upon their printing or processing speed, may be generally categorized as slow, medium, and high speed processors. In the case of slow processors like a fax machine, or simple graphic plotter, the document is propelled at less than 1 m/sec. The motion of this type of document may be easily modeled under reverse spaghetti problem category. In medium speed processors like commonly used copy machines, the typical document velocity may range from 1 to 3 or 4 m/s. In both these cases, commonly used bond papers are 8½" x 11" with mass density of about 60 to 100 gm/m^2, thickness of each sheet in the range of 0.01 to 0.02 mm and weight stiffness ratio of about 50 to 100.

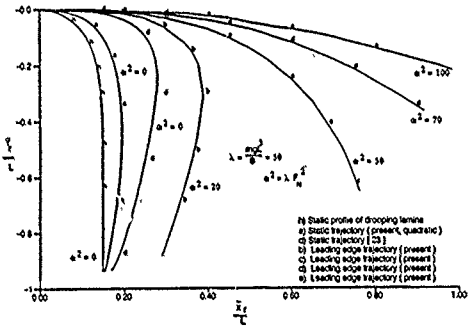


Figure 3. Reverse spaghetti problem showing the leading edge trajectory for various nondimensional velocities.

Figure 3 shows the effect of various nondimensional velocities for a specific value of weight stiffness ratio. As no account of aerodynamic forces are considered, results of this

analysis gives a good prédiction of the document performance for low to lower medium velocities. As in static case, the quadratic element solution better predicts the document kinematics. The series solution given by Mansfield and Simmonds [2] is also produced as a part of the overall solution scheme for comparison purposes. In fact, the series solution serves as a good starting solution to the FEM approximation. In the program, when the user chooses to run this case, then the program automatically calculates the series solution for each time step before proceeding with finite element solution. The user has an option to indicate for how many time steps this series solution is to be used as a starting solution for iterations within each time step. However, for higher nondimensional velocity ratios unless a smaller time interval ($\Delta\tau$) is chosen this procedure leads to convergence problems.

REFERENCES

[1] Wickert, J. A. and Mote, C. D., Jr., "Current Research on the Vibration and Stability of Axially Moving Materials," EUROMECH 223, Colloquium on Vibration and Stability of Axially-Moving Materials, TAMPERE, Finland, June 16-18, 1987.

[2] Carrier, G. F., "The Spaghetti Problem," American Mathematics Monthly, Vol. 56, pp. 669-672, 1949.

[3] Mansfield, L., Simmonds, J. G., "The Reverse Spaghetti Problem: Drooping Motion of an Elastica Issuing from a Horizontal Guide," ASME Journal of Applied Mechanics, Vol. 54, pp. 147-150, 1987.

INTERFACIAL HEAT TRANSFER BY EVERSION

Winston Khan
University of Puerto Rico
Mayaguez, Puerto Rico

Abstract: Heat transfer techniques are well established with an abundance of relevant literature. However, this paper attempts to add a different perspective through the process of convective eversion, which implies the mixing of fluid layers hot or cold, and, in this particular case the mixing of cold turbulent bulk fluid with hot interfacial fluid through the process of surface renewal, which was first propounded by Danckwerts in mass transfer processes.

I. INTRODUCTION

Danckwerts theory of surface renewal was propounded to explain mass transfer across interfaces where turbulence from the bulk phase extended to such interfaces, and where stagnant layer theory led to erroneous results.

The theory of Danckwerts not only led to more satisfactory results for mass transfer, but was more physically realistic in explaining the mechanism by which transfer took place, and, in essence described the more modern view of convective eversion.

Convective eversion is really the mixing of adjacent layers or regions of fluid, with an interchange of all physical properties, such as mass, heat, momentum, and energy. This notion can easily be extended to embrace the mixing of turbulent bulk fluid with interfacial fluid by the propagation of eddies or migration of eddies into the interface, which Danckwerts described as turbulence extending right into the interface due to its very nature of a mass of eddies, which incessantly change shapes and sizes and fluctuate randomly beneath the surface.

It is in this sense that the principle of convective eversion is applied, that is, the mixing of propelled eddies from the bulk into the interface, which bring their own physical properties and displace surface elements with corresponding physical properties, but different values.

II. HIGHLIGHTS OF DANCKWERTS THEORY

(Adapted to Heat Transfer)

When a liquid is in turbulent motion it is a mass of eddies which incessantly change their conformation and position. These eddies are pictured as continually exposing fresh surfaces to the gas or other phase, while sweeping away and mixing into the bulk parts of the surface which have been in contact with the other phase for varying lengths of time. The assumption is made that during the time of residence of any portion of the liquid it absorbs heat at a rate given by

$$\Psi(\theta) = (T^x - T)\sqrt{\frac{k}{\pi\theta}} \qquad (1)$$

which assumes that the scale of turbulence is much greater than the depth of penetration of heat conduction from the surface. Here, T^x is the interface temperature, T is the bulk fluid temperature, k is the thermal diffusivity and θ is the exposure time of eddy to upper phase.

In particular, consider a liquid which is maintained in turbulent motion by stirring at a steady rate. The total area of the surface exposed to the other phase being unity and the average rate of heat absorption is uniform over the area. The motion of the liquid will continually replace with fresh surface elements that are older and exposed for a finite time interval. The mean rate of production of fresh surface will be constant and equal to S and the chance of an element of surface being replaced within a given time is assumed to be independent of its age: then the fractional rate of replacement of elements in any age group is equal to S.

Let the area of surface comprising those elements having ages between θ and $(\theta + d\theta)$ by $\phi(\theta)d\theta$. It is not difficult to show that

$$\int_0^\infty \phi\, d\theta = 1 \quad \text{and} \quad \phi = S\, e^{-S\theta}. \qquad (2)$$

The rate of absorption of heat into those elements of surface having age θ and combined area $Se^{-S\theta}$ is obtained from above as,

$$R = (T^x - T)\, Se^{-S\theta}\sqrt{\frac{k}{\pi\theta}}\, d\theta \qquad (3)$$

After some mathematical manipulation we find that

$$R = (T^x - T)\sqrt{kS}\,. \qquad (4)$$

III. EXTENSION OF CONVECTIVE EVERSION TO CONTAMINATED SURFACES

In the physical model contemplated we envisage eddies as pieces of fluid impelled into the interface from the bulk turbulent phase. We also interpret S in the model, which is the fractional rate of replacement of fresh liquid in the interface, as the eddy frequency.

Interfacial heat transfer by convective eversion should be damped and sensitive to small quantities of surface impurities, and, this damping must be hydrodynamic in nature, hence S the eddy frequency must be damped.

Hence S must be a function of C_s^{-1}, the surface compressional modulus of elasticity of the film molecules.

Let V_0 be a characteristic velocity with which an eddy is impelled into the interface. The eddy possesses kinetic energy proportional to $\rho\lambda^3 V_0^2$, where ρ is the density of the liquid and λ the average size of an eddy.

This energy is then expended at the interface in overcoming the forces of surface tension and the surface compressional modulus of elasticity, which behaves like surface tension but acts tangentially to the surface only.

From the consideration of energy balance, neglecting gravity, we get

$$\rho\lambda^3 V_0^2 \propto (C_s^{-1} + \gamma)\,\lambda^2$$

$$\text{or,} \quad \lambda \propto \frac{C_s^{-1} + \lambda}{\rho V_0^2} \qquad (5)$$

Since turbulence is characterized by many superimposed periodicities, then by analogy with wave motion we write $s = V_0\Lambda$, where the size of the eddy corresponds to the wavelength. Therefore, $s = \rho V_0^3/\lambda + C_s^{-1}$, revealing the damping effect of the surfactant in general as

$$R\propto (T^{\infty} - T)\left[\frac{\rho k v_0^3}{\lambda + C_s^{-1}}\right]^{1/2} \qquad (6)$$

For stirred systems $V_0 = NL$, where N is the stirring speed and L the size of the stirrer blades. Therefore,

$$R\propto (T^{\infty} - T)\frac{(R_e)^{3/2}}{(\gamma + C_s^{-1})^{1/2}} . \qquad (7)$$

For turbulence in general

$$R\propto \frac{(v_0)^{3/2}}{(\gamma + C_s^{-1})^{1/2}} (T^{\infty} - T) \qquad (8)$$

REFERENCES

1. P. V. Danckwerts, Significance of liquid film coefficients in gas absorption. Ind. Chem. Eng. 43, 1960-67 (1951).
2. J. T. Davies and W. Khan, Surface clearing of eddies. Chem. Engr. Sci. 20, 713-718 (1965).
3. J. T. Davies, Turbulence Phenomena, 1st Ed., Academic Press, London and New York (1972). W. Khan, pp. 189, 224, 229, 248, 251, 258, 259, 272.
4. G. S. H. Lock and S. Park, A numerical model of convective eversion. Mathematical and Computer Modeling, Vol. 13, No. 7, 1990, pp. 107-116, Pergamon Press.

2-D (DIFFERENTIAL-DELAY) IMPLICIT-SYSTEMS

Stephen L. Campbell*
Department of Mathematics
& Center for Research in Scientific Computing
North Carolina State University
Raleigh, NC 27695-8205 USA

Abstract. This paper discusses linear time invariant differential-delay descriptor systems. The literature is surveyed and several observations are made.

1. INTRODUCTION

The general linear time invariant 2-D system takes the form

$$AD_tD_sx(t,s) + BD_tx(t,s) + CD_sx(t,s) + Dx(t,s) = Eu(t,s) \quad (1)$$

where $x(t,s)$ is a function of two variables and $\mathcal{D}_t$ is either a differential operator d/dr or a shift (delay) operator $\mathcal{D}_tx(\tau) = x(\tau - \alpha)$. There is an extensive literature on the partial differential equations, $Ax_{ts} + Bx_t + Cx_s + Dx = Eu$ where both operators are differentiation although the singular case is less studied. There is a growing literature on the classical 2-D system [3],[9] when both operators are discrete

$$Ax_{i,j} + Bx_{i,j+1} + Cx_{i+1,j} + Dx_{i+1,j+1} = Eu_{i+1,j+1} \quad (2)$$

Less work has been done on implicit singular differential delay systems where one operator is discrete and the other is continuous

$$Ax'(t-\alpha) + Bx'(t) + Cx(t-\alpha) + Dx(t) = Eu(t) \quad (3)$$

This is surprising given the importance of delay systems and the growing literature on singular systems [2],[4],[5]. The only paper we are aware of that deals directly with (3) is [3] which concerned the system

$$Bx'(t) + Cx(t-\alpha) + Dx(t) = Eu(t) \quad (4)$$

In this paper we make several observations about (3). The need for brevity prevents a careful discussion of the applications of delay equations. However, one example is instructive.

Example 1 Suppose we have a descriptor control system

$$Bx'(t) + Dx(t) = Eu(t) \quad (5a)$$
$$y(t) = Cx(t) \quad (5b)$$

and implement a feedback control law $u(t) = Ky(t) + v(t)$. However, α time units elapse between the observation $y(t)$ and application of the control. Then (5b) is $Bx'(t) + Dx(t) = Eu(t-\alpha)$. The closed loop system is then in the form (4)

$$Bx'(t) - EKCx(t-\alpha) + Dx(t) = Ev(t) \quad (6)$$

If we feed back velocity information using $u(t) = Ky(t) + Ly'(t) + v(t)$ the system is in the form (3)

$$-ELCx'(t-\alpha) + Bx'(t) - EKCx(t-\alpha) + Dx(t) = Ev(t) \quad (7)$$

Derivative feedback can sometimes regularize a 1-D singular system. Equation (7) shows that derivative feedback with delay can never regularize the system by producing a nonsingular A coefficient.

With (2) any result based on the pencil $\lambda D + B$ implied a similar result based on the pencil $\lambda D + C$. The differences between the differential and delay operators problem makes such a duality much less obvious in the differential-delay case.

*Research supported in part by the U.S. Army Research Office under DAAL03-89-D-0003, and the National Science Foundation under ECS-9012909.

2. NOTATION

Let α be a nonnegative real number and $x_1(t)$ a function defined on the t-interval $[0,\alpha]$. The solution $x(t)$ of (3) is $x(t) = x_i(t_\alpha)$ for $t \in [(i-1)\alpha, i\alpha]$ for $i \geq 0$, where

$$m_\alpha = \max\{\text{integers } m : t - m\alpha \leq 0\} \quad (8a)$$
$$t_\alpha = t - (m_\alpha - 1)\alpha \quad (8b)$$

We are interested in (3) for $t \geq 0$. The function $x_0(t)$ gives the initial data for (3). With this notation (3) becomes

$$Ax_i'(t) + Bx_{i+1}'(t) + Cx_i(t) + Dx_{i+1}(t) = Eu_{i+1}(t), \quad 0 < t \leq \alpha, \ i \geq 0 \quad (9)$$

3. NONSINGULAR DELAY EQUATIONS

The scalar delay equation

$$ax'(t-\alpha) + bx'(t) + cx(t-\alpha) + dx(t) = f(t) \quad (10)$$

is: of *retarded type* if $a = 0, b \neq 0$, of *neutral type* if $a \neq 0, b \neq 0$, and of *advanced type* if $a \neq 0, b = 0$ and $d \neq 0$ [1],[8]

Suppose we have the delay equation of retarded type

$$x_{i+1}'(t) + dx_{i+1}(t) + cx_i(t) = Eu_{i+1}(t), \quad 0 < t \leq \alpha, \ i \geq 0 \quad (11)$$

Given $x_0(t)$, we may recursively solve (11) by taking the initial conditions as $x_{i+1}(0) = x_i(\alpha)$, $i \geq 0$ If $u(t)$ is infinitely differentiable, the solution will be continuous and not differentiable at $t = 0$ if $x_0'(\alpha^-) \neq x_1'(0^+)$ Succeeding $x_i(t)$ become smoother, and $x(t)$ will be at least $i-1$ times continuously differentiable at $t = i\alpha$. This is important for numerical methods since a point of reduced smoothness for $x(t)$ can create difficulties with convergence and error control. If $u(t)$ has a jump discontinuity at $t = c$, then $x(t)$ will still be $i-1$ times differentiable at $t = c + i\alpha$.

For an equation of advanced type, (10) gives $x(t) = d^{-1}[f(t) - ax'(t-\alpha) - cx(t-\alpha)]$ so that there is no smoothing effect

4. $\lambda B + D$ REGULAR

By letting $y(t) = x(t-\alpha)$, $z = [x^T, y^T]^T$, (3) can be written

$$\begin{bmatrix} B & A \\ 0 & 0 \end{bmatrix} z'(t) + \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix} z(t) + \begin{bmatrix} C & 0 \\ -I & 0 \end{bmatrix} z(t-\alpha) = \begin{bmatrix} E \\ 0 \end{bmatrix} u(t) \quad (12)$$

which is (3) with $A = 0$. Thus a singular differential-delay system which appears retarded may actually include subsystems of all types

In this note we shall consider systems (3) for which $\lambda B + D$ is a regular pencil. The corresponding pencil for (12) is

$$\begin{bmatrix} \lambda B + D & A \\ 0 & I \end{bmatrix} \quad (13)$$

Proposition 1 *The pencil* (13) *is regular if and only if* $\lambda B + D$ *is regular. Also, the index of* (13) *is the same as the index of the pencil* $\lambda B + D$.

Suppose that $\lambda B + D$ is a regular pencil. By Proposition 1 and (12) we may assume that $A = 0$ and the system is in the form (4). Then by the usual singular system theory [4],[5] we may solve (9) for $x_{i+1}(t)$ given $x_i(t)$. Since we are interested in behavior more than solution formula here, we will perform constant coordinate changes on (4) based on the pencil $\lambda B + D$ to get

$$\begin{aligned} z'(t) + D_1 z(t) &= E_1 u(t) + C_1 z(t-\alpha) + C_2 w(t-\alpha) \quad (14a) \\ N w'(t) + w(t) &= E_2 u(t) + C_3 z(t-\alpha) + C_4 w(t-\alpha) \quad (14b) \end{aligned}$$

where $N^k = 0, N^{k-1} \neq 0$ for some integer $k \geq 1$ [4],[5]. Then

$$w(t) = \sum_{i=0}^{k-1} (-N)^i [E_2 u^{(i)}(t) + C_3 z^{(i)}(t-\alpha) + C_4 w^{(i)}(t-\alpha)] \quad (14c)$$

Equations (14b), (14c) have several consequences. Solutions of (3) can be continuous on only finite intervals even if $u(t)$ is infinitely differentiable. An example is given in [3] where if C_n is the set of initial conditions whose solutions are continuous on $[0, n]$, then C_{n+1} is a proper subset of C_n for every $n \geq 0$. Secondly, discontinuities do not necessarily get smoothed out as with the nonsingular problem.

Example 2 Consider the 2-D system with $\alpha = 1$, $z = [x^T, y^T]^T$.

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} z'(t) + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} z(t) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} z(t-1) + \begin{bmatrix} 0 \\ v(t) \end{bmatrix} \quad (15)$$

The x component of z satisfies the equation of advanced type $x(t) = x'(t-1) + v'(t)$ so that $x(t) = x^{(m)}(t-m) + \sum_{i=0}^{m-1} v^{(i+1)}(t-i)$ Note that $x(t)$ becomes progressively less smooth. Discontinuities of $v(t)$ not only persist, but become more impulsive.

Index One Systems

For the remainder of this subsection assume the system is (4) and that $\lambda B + D$ is an index one pencil If $B = 0$, then D is nonsingular, and (14) becomes the functional equation

$$w(t) = C_4 w(t-\alpha) + E_2 u(t) \quad (16)$$

with solution

$$w(t) = C_4^{m_\alpha} w(t - m_\alpha \alpha) + \sum_{i=0}^{m_\alpha - 1} C_4^i E_2 u(t - i\alpha) \quad (17)$$

where m_α is given by (8). If $B \neq 0$ and B is singular, (14) is

$$\begin{aligned} z'(t) + D_1 z(t) &= E_1 u(t) + C_1 z(t-\alpha) + C_2 w(t-\alpha) \quad (18a) \\ w(t) &= E_2 u(t) + C_3 z(t-\alpha) + C_4 w(t-\alpha) \quad (18b) \end{aligned}$$

Given $z(t), w(t)$ on $(i\alpha, (i+1)\alpha]$, the system (18) gives $z(t), w(t)$ on $((i+1)\alpha, (i+2)\alpha]$. Discontinuities interior to the interval may be carried over in $w(t)$ but not in $z(t)$. If $u(t)$ is piecewise continuous, we will have that $w(t)$ is piecewise continuous but $z(t)$ is continuous. If $u(t)$ is continuous, and the initial conditions are continuous, then $w(t)$ is continuous except possibly at integer multiples of α and $z(t)$ will be differentiable except possibly at integer multiples of α where $z(t)$ will be only continuous. However, $z(t)$ will have left and right hand derivatives at these points.

Thus if $\lambda B + D$ is index one, (3) acts like either a neutral or a retarded system. If $\lambda B + D$ is higher index, the behavior will possibly be of advanced type.

5. CONTINUITY IN α

If we set $\alpha = 0$ in the system (3), we get

$$(A + B)x'(t) + (C + D)x(t) = Eu(t) \quad (19)$$

The pencil $\lambda(A+B)+(C+D)$ need not be regular if $\lambda B+D$ is. Even if A is nonsingular, (19) can be singular and the pencil $\lambda(A+B)+(C+D)$ may be regular or singular [10]. Do the solutions converge as $\alpha \to 0^+$?

Suppose that we have (4). One likely necessary condition is that the roots of the complex function

$$g(s) = \det(sB + D + Ce^{-s\alpha}) \quad (20)$$

are bounded above in real part as $\alpha \to 0^+$. See [1],[8] for the mathematical techniques. Here we look at a simple index one example.

Example 3 Consider the index one system

$$w(t) = rw(t-\alpha) + v(t) \quad (21)$$

whose solution is

$$w_\alpha(t) = r^m w(t - m_\alpha \alpha) + \sum_{i=0}^{m_\alpha - 1} r^i v(t - i\alpha) \quad (22)$$

The limiting equation for (21) as $\alpha \to 0^+$ is

$$w(t) = rw(t) + v(t) \quad (23)$$

which is degenerate if $r = 1$ and for $r \neq 1$ has the unique solution

$$w(t) = \frac{1}{1-r} v(t) \quad (24)$$

If $|r| > 1$, and either $v(t) \not\equiv 0$ or $w(0^-) \neq 0$, then $w_\alpha(t)$ does not converge to (24) On the other hand, if $|r| < 1$, and $v(t)$ is continuous on the interval $(-1, t]$, then $w_\alpha(t)$ does converge to (24) as $\alpha \to 0^+$. Note that for (21) we have that

$$g(s) = \det(sB + D + Ce^{-s\alpha}) = (1 - re^{-s\alpha}) \quad (25)$$

which has the roots $s = -\frac{1}{\alpha}(-\ln|r| + i2n\pi)$, $i = 0, 1, 2, \ldots$ The roots of (25) have real part bounded above as $\alpha \to 0^+$ if and only if $|r| \leq 1$ The case $r = 1$ shows that additional conditions beyond the boundedness of real parts are needed to get convergence as $\alpha \to 0^+$.

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, 1973.

[2] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Elsevier, 1989.

[3] S. L. Campbell, *Singular linear systems of differential equations with delays*, Applicable Analysis, II (1980), 129-136.

[4] S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, 1980.

[5] S. L Campbell, *Singular Systems of Differential Equations II*, Pitman, 1982.

[6] S. L. Campbell, *Comments on 2-D descriptor systems*, Automatica, 27 (1991), 189-192.

[7] S. L Campbell and C D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover Press reprint in press, 1991.

[8] J. Hale, *Functional Differential Equations*, Springer, 1971.

[9] T. Kaczorek, *Two-Dimensional Linear Systems*, Springer, 1985

[10] B. S. Mordukhovich, *Controllability, observability, and optimality in hereditary systems of neutral type*, Proc. 29th IEEE Conf. Decision & Control, 1990, 2960-2965.

ALGEBRAIC ASPECTS OF 2D SINGULAR SYSTEMS

E. Fornasini and S. Zampieri
Dept. of Electr. and Inform. University of Padova
via Gradenigo 6/a, 35131 Padova, Italy, fax 39-49-8287699

Abstract The paper investigates the behaviour $\mathcal{B}$ of a singular 2D system on a half plane. Some connections between the matrices appearing in the updating equations and the restrictions of $\mathcal{B}$ to the separation sets are presented.

1 Introduction

Consider a 2D system given by the following equation

$$\overline{E}x(h+1,k+1) = \overline{A}x(h,k+1) + \overline{B}x(h+1,k) \tag{1}$$

where $x \in \mathbb{R}^n$ and $\overline{E}, \overline{A}, \overline{B}$ are $q \times n$ matrices with entries in $\mathbb{R}$.

Clearly, if $\operatorname{rank}\overline{E} = n$, (1) can be reduced to the equation of an unforced nonsingular 2D system [1], as follows

$$\begin{aligned} x(h+1,k+1) \;=\; & (\overline{E}^T\overline{E})^{-1}\overline{E}^T\overline{A}x(h,k+1) \\ & +(\overline{E}^T\overline{E})^{-1}\overline{E}^T\overline{B}x(h+1,k) \end{aligned} \tag{2}$$

If $\operatorname{rank}\overline{E} = r < n$, we are allowed to introduce two nonsingular matrices $Q \in \mathbb{R}^{q\times q}$ and $N \in \mathbb{R}^{n\times n}$, such that

$$Q\overline{E}N = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \tag{3}$$

So, letting

$$x = N^{-1}x,\; Q\overline{A}N = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},\; Q\overline{B}N = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{4}$$

equation (1) can be rewritten as follows

$$\begin{aligned} x_1(h+1,k+1) = {} & A_{11}x_1(h,k+1) + B_{11}x_1(h+1,k) \\ & + A_{12}x_2(h,k+1) + B_{12}x_2(h+1,k) \\ 0 = {} & A_{21}x_1(h,k+1) + B_{21}x_1(h+1,k) \\ & + A_{22}x_2(h,k+1) + B_{22}x_2(h+1,k) \end{aligned} \tag{5}$$

In the particular case when $A_{21}, B_{21}, A_{22}, B_{22}$ are simultaneously zero, x_2 can be viewed as an $n-r$ dimensional input and (5) provides the state updating equation of a nonsingular 2D system. More generally, however, x_2 is the direct sum of exhogenous variables (i.e. inputs), and auxiliary variables that induce some dynamical constraints on the system trajectories, and (5) can be considered a singular 2D system, as studied in [2].

This paper constitutes a preliminary report on a research still in progress, concerning the analytical structure of the trajectories of system (5) in the half plane $\mathcal{H} = \{(h,k) : h+k \geq 0\}$. No "a priori" assumption is made on which components of x_2 can be given the role of exhogenous variables. Following the philosophy that underlies the behavioural approach by J. Willems and P. Rocha [3-5], the nature of the input functions is determined "a posteriori", after establishing what variables are constrained by equations (5).

2 An algebraic approach via duality

All signals x that will be considered in this paper are sequences indexed on the half plane $\mathcal{H}$ and taking values in some finite dimensional $\mathbb{R}$-vector space $x : \mathcal{H} \longrightarrow \mathbb{R}^n : (h,k) \longmapsto x(h,k)$. The single step updating structure (5) makes it convenient to introduce a partition of $\mathcal{H}$ into a countable family of separation sets $\mathcal{S}^i = \{(h,k) : h+k = i\}$, $i = 0,1,\dots$ and to associate with x a formal power series

$$X = \sum_{i=0}^{+\infty}\sum_{j=-\infty}^{+\infty} x(i+j,i-j)\xi^{-j}\lambda^{-i} \tag{6}$$

So, the "bilateral" formal power series $X^i = \sum_{j=-\infty}^{+\infty} x(i+j,i-j)\xi^{-j}$, $i = 0,1,\dots$ are associated with the restrictions of the signal x to the separation sets $\mathcal{S}^i$, $i = 0,1,\dots$.

Let denote by F^n and G^n respectively the spaces of polynomials in ξ, ξ^{-1}, λ and of formal power series in $\xi, \xi^{-1}, \lambda^{-1}$, with coefficients in $\mathbb{R}^n$. Introduce in $F^n \times G^n$ a nondegenerate bilinear function $\langle \cdot,\cdot \rangle_n$ that associates with a polynomial $p = \sum_{i=0}^{\ell}\sum_{j=-m}^{m} p_{ij}\xi^j\lambda^i$ in F^n and a series $X = \sum_{i=0}^{+\infty}\sum_{j=-\infty}^{+\infty} x(i+j,i-j)\xi^{-j}\lambda^{-i}$ in G^n the coefficient of the constant term in the Cauchy product p^TX

$$\langle p, X\rangle_n = \sum_{i=0}^{\ell}\sum_{j=-m}^{m} p_{ij}^T x(i,j). \tag{7}$$

Every series X in G^q induces a linear function φ_X on F^n, defined by $\varphi_X : p \longmapsto \langle p, X\rangle_n$. Moreover, the linear mapping that associates X with the linear function φ_X is an isomorphism of G^n onto the space $\mathcal{L}[F^n]$ of linear functions on F^n and, consequently, each series in G^n (or, equivalently, each signal $x : \mathcal{H} \to \mathbb{R}^n$) can be identified with an element of the algebraic dual space $\mathcal{L}[F^n]$. This accounts for the possibility of expressing many features of signal spaces with support in $\mathcal{H}$ in terms of properties of suitable subspaces of F^n.

Let $M(\xi,\xi^{-1},\lambda)$ be a $q \times n$ matrix with entries in $\mathbb{R}[\xi,\xi^{-1},\lambda]$ and consider the linear mappings

$$\begin{aligned} \mu : \; & F^q \longrightarrow F^n \; : \; p \longmapsto M^T(\xi,\xi^{-1},\lambda)p \\ \mu^* : \; & G^n \longrightarrow G^q \; : \; X \longmapsto M(\xi,\xi^{-1},\sigma)X \end{aligned} \tag{8}$$

Here σ is the shift operator in G^1

$$\sum w(i+j,i-j)\xi^{-j}\lambda^{-i} \xrightarrow{\sigma} \sum w(i+1+j,i+1-j)\xi^{-j}\lambda^{-i} \tag{9}$$

and μ and μ^* are dual mappings [5], as $\langle \mu p, X\rangle_n = \langle p, \mu^*X\rangle_q$ holds for all X in G^n and p in F^q. We therefore have

$$\ker\mu^* = (\operatorname{im}\mu)^\perp, \tag{10}$$

where $\operatorname{im}\mu$ denotes the $\mathbb{R}[\xi,\xi^{-1},\lambda]$-module generated by the columns of the matrix $M^T(\xi,\xi^{-1},\lambda)$.

In order to analyze the trajectories of system (5), we introduce the following series

$$X_\ell \;=\; \sum_{i=0}^{+\infty}\sum_{j=-\infty}^{+\infty} x_\ell(i+j,i-j)\xi^{-j}\lambda^{-i}, \quad \ell = 1,2 \tag{11}$$

and the matrix

$$M(\xi,\sigma) = \begin{bmatrix} \sigma I - A_{11} - B_{11}\xi & -A_{12} - B_{12}\xi \\ -A_{21} - B_{21}\xi & -A_{22} - B_{22}\xi \end{bmatrix} \tag{12}$$

The constraints induced on x by equation (5) are expressed as

$$M(\xi,\sigma)\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0$$

Therefore the behaviour of (5) can be viewed as the kernel of the linear operator μ^* or, alternatively, as the orthogonal subspace to the $\mathbb{R}[\xi,\xi^{-1},\lambda]$-module M generated by the columns of the matrix $M^T(\xi,\lambda)$:

$$\mathcal{B} = \left\{ \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in G^n \quad M(\xi,\sigma)\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0 \right\} = \ker\mu^* = M^\perp \tag{13}$$

In our context an important consequence stems directly from the fact that G^n is the algebraic dual $\mathcal{L}[F^n]$, namely

$$\mathcal{B}^\perp = (M^\perp)^\perp = M \tag{14}$$

Actually (14) shows that the module $\mathcal{M}$ is uniquely determined by $\mathcal{B}$, so that $\mathcal{B}$ can be described as the kernel of some matrix $\overline{M}(\xi,\sigma)$ if and only if the columns of both $M^T(\xi,\lambda)$ and $\overline{M}^T(\xi,\lambda)$ generate the same $\mathbb{R}[\xi,\xi^{-1},\lambda]$-module.

The duality theory provides also an useful tool for analyzing the restrictions $\mathcal{B}^{[0,k]}$ of the behaviour $\mathcal{B}$ to the sets $S^0 \cup S^1 \cup \cdots \cup S^k$. This is easily seen by considering the linear mappings

$$\begin{array}{ccccc} F_k^n & \xrightarrow{i} & F^n & \xrightarrow{\overline{\pi}} & F^n/\mathrm{im}\mu \\ G^n/\sigma^k G^n & \xleftarrow{\pi} & G^n & \xleftarrow{\overline{i}} & \ker\mu^* \end{array} \tag{15}$$

where F_k^n is the $\mathbb{R}[\xi,\xi^{-1}]$-submodule of the polynomial columns in F^n having degree less than or equal to k in the indeterminate λ, $G^n/\sigma^k G^n$ is (isomorphic to) the $\mathbb{R}[\xi,\xi^{-1}]$-submodule obtained by truncating in each series of G^n all terms with degree greater than k w.r. to λ^{-1}, the maps i and $\overline{i}$ are canonical injections, π and $\overline{\pi}$ are canonical projections.

Obviously $G^n/\sigma^k G^n$ is isomorphic with the space $\mathcal{L}[F_k^n]$ of linear functions on F_k^n. Moreover $F^n/\mathrm{im}\mu$ is isomorphic with a direct complement of $\mathrm{im}\mu$ in F^n, and using the duality theory on direct decompositions [6] gives $\ker\mu^* = (\mathrm{im}\mu)^\perp \cong \mathcal{L}[F^n/\mathrm{im}\mu]$. The first and the last space on the second row of (15) can be viewed as the algebraic duals of the corresponding spaces on the first row and the maps $\overline{\pi}\circ i$, $\pi\circ\overline{i}$ in (15) are dual linear maps w r. to the bilinear function induced on the pairs $(F_k^n,\ G^n/\sigma^k G^n)$ and $(F^n/\mathrm{im}\mu,\ \ker\mu^*)$. Consequently the restriction $\mathcal{B}^{[0,k]}$ is given by

$$\mathcal{B}^{[0,k]} \cong \mathrm{im}(\pi\circ\overline{i}) = \ker(\overline{\pi}\circ i)^\perp \tag{16}$$

The above relation characterizes $\mathcal{B}^{[0,k]}$ as the subspace of all signals with support in $S^0 \cup S^1 \cup \cdots \cup S^k$ and values in $\mathbb{R}^n$ that correspond to formal power series $\sum_{i=0}^k X^i\lambda^{-i}$ satisfying the orthogonality condition

$$\Big\langle \sum_{i=0}^k c_i(\xi)\lambda^i\ ,\ \sum_{i=0}^k X^i\lambda^{-i} \Big\rangle_n = [c_0(\xi)\ c_1(\xi)\ \cdots\ c_k(\xi)] \begin{bmatrix} X^0 \\ X^1 \\ \vdots \\ X^k \end{bmatrix} = 0 \tag{17}$$

for all polynomial vectors $\sum_{i=0}^k c_i(\xi)\lambda^i$ in the $\mathbb{R}[\xi,\xi^{-1},\lambda]$-module $\mathrm{im}\mu$.

The $\mathbb{R}[\xi,\xi^{-1}]$-submodule of $\mathbb{R}^{1\times n(k+1)}[\xi,\xi^{-1}]$ whose elements are the rows $[c_0(\xi)\ c_1(\xi)\ \cdots\ c_k(\xi)]$ that satisfy the condition $\sum_{i=0}^k c_i(\xi)\lambda^i \in \mathrm{im}\mu$ is finitely generated. Therefore there exists a polynomial matrix $C^{[0,k]}(\xi)$ with $n(k+1)$ columns such that $\mathcal{B}^{[0,k]} = \ker C^{[0,k]}(\xi)$.

In the next section we shall take advantage of the particular structure of $\mathcal{M}(\xi,\sigma)$ given by (12), when determing the $\mathbb{R}[\xi,\xi^{-1}]$-submodule $\mathcal{B}^{[0,k]}$.

3 Computation of trajectories

The following lemma directly provides a matrix $C^{[0,1]}(\xi)$ whose rows are given in terms of submatrices A_{ij} and B_{ij} that appear in the partition (12) The proof is based on Cayley-Hamilton theorem and can be found in [7].

Lemma Let A_{ij} and B_{ij} be as in (12) and define the polynomial matrices:

$$A_{ij} := A_{ij} + B_{ij}\xi, \quad i,j = 1,2$$

$$[C_0(\xi)\quad C_1(\xi)] = \left[\begin{array}{cc|cc} A_{21} & A_{22} & 0 & 0 \\ A_{11} & A_{12} & I_r & 0 \\ 0 & 0 & A_{21} & A_{22} \\ 0 & 0 & A_{21}A_{11} & A_{21}A_{12} \\ 0 & 0 & A_{21}A_{11}^2 & A_{21}A_{11}A_{12} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & A_{21}A_{11}^r & A_{21}A_{11}^{r-1}A_{12} \end{array}\right]$$

Then $C^{[0,1]}(\xi) = [C_0(\xi) \mid C_1(\xi)]$ and $\mathcal{B}^{[0,1]} = \ker C^{[0,1]}(\xi)$ or, equivalently,

$$\begin{bmatrix} X^0 \\ X^1 \end{bmatrix} \in \mathcal{B}^{[0,1]} \Leftrightarrow C_0X^0 = -C_1X^1 \tag{18}$$

Premultiplying both C_0 and C_1 by a suitable unimodular matrix U, one gets

$$UC_0 = \begin{bmatrix} D_0' \\ D_0 \\ 0 \end{bmatrix}, \quad -UC_1 = \begin{bmatrix} D_1' \\ 0 \\ 0 \end{bmatrix} \tag{19}$$

where both D_0 and D_1 have full row rank. Just rewriting (18) as

$$\begin{aligned} D_0X^0 &= 0 \\ D_1X^1 &= D_0'X^0 \end{aligned} \tag{20}$$

we easily see that all solutions of equation (20.1) can be viewed as restrictionsof admissible trajectories to the separation set S^0. In fact D_1 has full row rank and, therefore, given any X^0, eq. (20.2) can be fulfilled by suitably chosen values of X^1.

We are now in a position for establishing the following

Theorem A signal $X = \sum_{i=0}^{+\infty} X^i\lambda^{-i}$ belongs to $\mathcal{B}$ if and only if X^i satisfy the following equations

$$\begin{aligned} &D_0X^0 = 0 \\ &\begin{bmatrix} D_1 \\ D_0 \end{bmatrix} X^{i+1} = \begin{bmatrix} D_0' \\ 0 \end{bmatrix} X^i, \quad i = 0,1,\ldots \end{aligned} \tag{21}$$

PROOF Suppose that X satisfies (21). Then we have

$$C^{[0,1]}(\xi)\begin{bmatrix} X^i \\ X^{i+1} \end{bmatrix} = 0, \quad i = 0,1,\ldots \tag{22}$$

which implies

$$X^i + \lambda^{-1}X^{i+1} \in \mathcal{B}^{[0,1]}, \quad i = 0,1,\ldots \tag{23}$$

The degree of all columns in $M^T(\xi,\lambda)$ w.r.to λ is less than or equal to one. So, any such column can be written as $c_0(\xi) + \lambda c_1(\xi)$ and we have

$$\begin{aligned} &\langle (c_0(\xi) + \lambda c_1(\xi))\lambda^i\xi^j, X\rangle_n \\ &= \langle (c_0(\xi) + \lambda c_1(\xi))\xi^j, X^i + \lambda^{-1}X^{i+1} + \ldots\rangle_n \\ &= \langle (c_0(\xi) + \lambda c_1(\xi))\xi^j, X^i + \lambda^{-1}X^{i+1}\rangle_n = 0 \end{aligned} \tag{24}$$

as a consequence of (16) and (23) This shows that X is orthogonal to $\mathrm{im}\mu$ and therefore $X \in \mathcal{B}$. The converse is obvious.

Equations (21) provide a recursive procedure for generating the system trajectories. Moreover, the difference $n - \mathrm{rank}\begin{bmatrix} D_1 \\ D_0 \end{bmatrix}$ gives the number of free variables that appear in system (5), i.e. the variables that can be arbitrarily chosen on all separation sets S^i.

4 References

1. E Fornasini, G.Marchesini *Doubly indexed dynamical systems. state space models and structural properties*, Math. Sys. Theory, 12, 123-29, 1978
2. T.Kaczorek *Singular multidimensional linear discrete systems*, Proc. 1st Nat. Symp. Aut. Robot., Athens, 71-91, 1987
3. J.C.Willems *Models for dynamics*, Dyn. Rep., 2, 171-269, 1989
4. P.Rocha, J.C.Willems *Canonical computational forms of AR 2D systems*, Multidim.Sys Sign.Proc., 1, 251-78, 1990
5. P.Rocha *Structure and representation of 2D systems*, Ph.D Thesis, Rijksuniversiteit Groningen, 1990
6. W.Greub *Linear Algebra*, Springer, 1975
7. E Fornasini, S Zampieri *Singular 2D systems. a behavioural approach*, submitted

Geometric Theory for 2-D Implicit Roesser Systems

F. L. Lewis and A. Karamancioğlu
Automation and Robotics Research Institute
University of Texas at Arlington

Abstract

The two-dimensional implicit Roesser model is considered from a geometric point of view. Its (A,E,B) and (E,A,B)-invariant subspaces are related to the existence of solutions within a certain subspace. The (A,E,B)-invariant subspace of the dual system is utilized in designing an asymptotic unknown-input observer. An illustrative example is included.

1 Introduction

The importance of the two-dimensional (2-D) implicit models is due to their multidirectional dynamic structure as well as their ability to express algebraic relationships among the system *semistates*. They can model systems with boundary conditions specified all around the domain of interest, such as systems modelled by the elliptic equation. This cannot be accomplished by the 2-D state-space models.

We consider the implicit Roesser model (IRM) [1] from a geometric point of view. This point of view allows one to represent infinitely many system behaviors in a very efficient way. It also allows one to characterize the boundary conditions which are projections of a solution onto the boundaries of the domain of interest.

In the next section we briefly present the semistate propagation mechanism of the IRM. In Section 3 we design an asymptotic observer which uses only the outputs, and not the inputs, of the actual system to reconstruct the semistate. We include an illustrative example.

2 Dynamics of the Implicit Roesser Model

Consider the implicit Roesser model

$$E\begin{bmatrix} x^h_{i+1,j} \\ x^v_{i,j+1} \end{bmatrix} = A\begin{bmatrix} x^h_{i,j} \\ x^v_{i,j} \end{bmatrix} + Bu_{i,j}, \qquad y_{i,j} = C\begin{bmatrix} x^h_{i+1,j} \\ x^v_{i,j+1} \end{bmatrix} \tag{1}$$

where $x^h \in \mathbb{R}^{n1}$, $x^v \in \mathbb{R}^{n2}$, and $u \in \mathbb{R}^m$ are the horizontal semistate, vertical semistate, and input vectors respectively. The constant coefficient matrices E, A, and B are of appropriate dimensions. We assume neither invertibility nor squareness of any of them. Therefore, solution to (1) may or may not exist.

Denote $\{0,1,\ldots,N-1\}$ by $\underline{N}$. We say a semistate sequence $(x_{i,j})$ is *a solution* if it satisfies (1) in the domain $\{(i,j) | i \in \underline{M} \text{ and } j \in \underline{N}\}$.

Define $E_1 := E$ diag$\{I_{n1},0\}$ and $E_2 := E$ diag$\{0,I_{n2}\}$, and A_i likewise. To characterize the semistate propagation of IRM we need to define certain invariant subspaces and some subspaces derived from them. We first consider the forward propagation case.

Define the subspace $\mathrm{V} \subset \mathbb{R}^n$ as *a (2-D) (A,E,B)-(controlled)-invariant subspace for (1)* if it satisfies

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}\mathrm{V} \subset \begin{bmatrix} E & 0 \\ 0 & E \end{bmatrix}\begin{bmatrix} \mathrm{V} \\ \mathrm{V} \end{bmatrix} + \mathrm{Im}\begin{bmatrix} B & 0 \\ 0 & B \end{bmatrix}. \tag{2}$$

The *horizontal* and *vertical (A,E,B)-invariant subspaces* of (1)

$$\mathrm{V}^h := \{x^h \in \mathbb{R}^{n1} | \begin{bmatrix} x^h \\ w \end{bmatrix} \in \mathrm{V} \text{ for some } w \in \mathbb{R}^{n2}\}$$

and

$$\mathrm{V}^v := \{x^v \in \mathbb{R}^{n2} | \begin{bmatrix} y \\ x^v \end{bmatrix} \in \mathrm{V} \text{ for some } y \in \mathbb{R}^{n1}\}$$

are useful in describing the forward propagation mechanism of the IRM:

Theorem 1. [5] Let V^h and V^v be a pair of horizontal and vertical (A,E,B)-invariant subspaces for (1). If $x^v_{i,0} \in \mathrm{V}^v \; \forall i \in \underline{M}$ and $x^h_{0,j} \in \mathrm{V}^h \; \forall j \in \underline{N}$ then there exist a solution for (1) such that the horizontal and vertical semistates of the solution are contained in V^h and V^v respectively. □

Using this theorem one can test each semistate on the left and bottom boundaries of the domain of interest to decide whether it gives rise to a solution.

For the backward propagation case define *a (2-D) (E,A,B)-(controlled)-invariant subspace* S for the IRM by replacing A_1, A_2, and E in (2) by E_1, E_2, and A respectively. Then the *horizontal* and *vertical (E,A,B)-invariant subspaces*

$$\mathrm{S}^h := \{x^h \in \mathbb{R}^{n1} | \begin{bmatrix} x^h \\ w \end{bmatrix} \in \mathrm{S} \text{ for some } w \in \mathbb{R}^{n2}\}$$

and

$$\mathrm{S}^v := \{x^v \in \mathbb{R}^{n2} | \begin{bmatrix} y \\ x^v \end{bmatrix} \in \mathrm{S} \text{ for some } y \in \mathbb{R}^{n1}\}$$

describe the backward propagation of the semistates, as formalized by the next theorem.

Theorem 2. [5] Let S^h and S^v be a pair of horizontal and vertical (E,A,B)-invariant subspaces for (1). If the boundary conditions $x^v_{i,N} \in \mathrm{S}^v \; \forall i \in \underline{M}$ and $x^h_{M,j} \in \mathrm{S}^h \; \forall j \in \underline{N}$ then there exists a solution for (1) with these boundary conditions such that each semistate in the solution is contained in S. □

Note that the conclusion of Theorem 2 is stronger than that of Theorem 1.

Theorem 3. [5] Let V^h, V^v, S^h, and S^v be defined as in the preceeding theorems. If the internal semistates $x^v_{i,k} \in \mathrm{S}^v \cap \mathrm{V}^v \; \forall i \in \underline{l}$ and $x^h_{l,j} \in \mathrm{S}^h \cap \mathrm{V}^h \; \forall j \in \underline{k}$ then there exists a solution with these internal semistates such that horizontal and vertical semistates are contained in $\mathrm{S}^h \cap \mathrm{V}^h$ and $\mathrm{S}^v \cap \mathrm{V}^v$ respectively. □

Research supported by NSF Grant ECS-8805932

3 Unknown-Input Observer Design

In this section we design an unknown-input observer for the IRM using geometric terms. The terminology *unknown-input* refers to the fact that the observer does not use the input u_{ij}, but only the outputs y_{ij} in reconstructing the semistate x_{ij}. Thus, u_{ij} could be an unknown disturbance acting on (1).

The existence of the observer will be formulated using the parameters of the dual system. By the *dual* system to (1) we mean the system with the coefficient matrices (E, A_1, A_2, B, C) replaced by $(E^T, A_1^T, A_2^T, C^T, B^T)$.

We call an (A,E,B)-invariant subspace for (1) contained in Ker C *an output nulling (A,E,B)-invariant subspace* for (1). Let V be an output-nulling (A,E,B)-invariant subspace of the dual system. Then using the Wonham techniques [6], one can show

$$\mathrm{T} \supset \begin{bmatrix} A_1 & A_2 & B \end{bmatrix} \left\{ \begin{bmatrix} E^{-1}\mathrm{T} \\ E^{-1}\mathrm{T} \\ \mathrm{R}^m \end{bmatrix} \cap \mathrm{Ker} \begin{bmatrix} C & 0 & 0 \\ 0 & C & 0 \end{bmatrix} \right\} \quad (3)$$

holds with T an orthogonal complement of V. We call T a *conditioned-invariant subspace* of (1) [2].

Select any Q of full row rank such that $\mathrm{Ker}Q = \mathrm{T}$. Then (3) is equivalent to the existence of matrices Γ_1, Γ_2, Λ_1, and Λ_2 such that the 2-D Lyapunov equation

$$Q \begin{bmatrix} A_1 & A_2 & B \end{bmatrix} = \begin{bmatrix} \Gamma_1 QE & \Gamma_2 QE & 0 \end{bmatrix} + \begin{bmatrix} \Lambda_1 C & \Lambda_2 C & 0 \end{bmatrix} \quad (4)$$

holds.

The Lyapunov equation (4) yields a 2-D *state-space* observer for (1) if the 2-D *state-space* triple (I, Γ_1, Γ_2) satisfying (4) is asymptotically stable in the standard 2-D state-variable sense [3]:

Theorem 4. [6] Let Q satisfy (4), so that $\mathrm{T} = \mathrm{Ker}(Q)$, with Q a full-row-rank solution to (4) for some Γ_1, Γ_2, Λ_1, and Λ_2. If the triple (I, Γ_1, Γ_2) is asymptotically stable, then the state-space system

$$z_{i+1,j+1} = \Gamma_1 z_{i,j+1} + \Gamma_2 z_{i+1,j} + \Lambda_1 y_{i,j+1} + \Lambda_2 y_{i+1,j}. \quad (5)$$

estimates QEx_{ij} with decreasing error as (i,j) moves away from the initial boundary. □

Note that if $\mathrm{T} = 0$ and the corresponding Lyapunov equation is stable, then the entire external variable Ex_{ij} can be reconstructed. Therefore, we may alternatively call T the *unknown-input unobservable subspace*. Note also that the observer is a 2-D *state-space* system, not an implicit system.

Example 1.

Consider the system modelled by

$$\begin{bmatrix} 5 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x^h_{i+1,j} \\ x^v_{i,j+1} \end{bmatrix} = \begin{bmatrix} 2 & 4 & 0 \\ 0 & 0 & .3 \\ 0 & .5 & 1 \end{bmatrix} \begin{bmatrix} x^h_{ij} \\ x^v_{ij} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u_{ij}$$

$$y = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x^h_{ij} \\ x^v_{ij} \end{bmatrix} \quad (6)$$

with $n_1 = 2$ and $n_2 = 2$. We wish to design an asymptotic observer for the states that are not seen through the output equation directly.

We use the *shifted representation* of Equation (6) (see [6])

$$E\tilde{x}_{i+1,j+1} = A_1 \tilde{x}_{i,j+1} + A_2 \tilde{x}_{i+1,j} + Bu_{i,j},$$

which is obtained by the transformation $\tilde{x}_{ij} = \begin{bmatrix} x^h_{i,j-1} \\ x^v_{i-1,j} \end{bmatrix}$. Observing that $\mathrm{span}\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$ is a conditioned-invariant subspace for (6), one may select

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

to observe $QE\tilde{x}_{ij} = \begin{bmatrix} 5\tilde{x}^{h1}_{ij} + 2\tilde{x}^{h2}_{ij} \\ 4\tilde{x}^{v}_{ij} \end{bmatrix}$. A solution to the Lyapunov equation yields the unknown-input state-space observer

$$z_{i+1,j+1} = \begin{bmatrix} .4 & 0 \\ 0 & 0 \end{bmatrix} z_{i,j+1} + \begin{bmatrix} 0 & 0 \\ 0 & .25 \end{bmatrix} z_{i+1,j} + \begin{bmatrix} 3.2 \\ .5 \end{bmatrix} y_{i,j+1}.$$

The error equation

$$\begin{aligned} e_{i+1,j+1} &:= z_{i+1,j+1} - QE\tilde{x}_{i+1,j+1} \\ &= \begin{bmatrix} .4 & 0 \\ 0 & 0 \end{bmatrix} e_{i,j+1} + \begin{bmatrix} 0 & 0 \\ 0 & .25 \end{bmatrix} e_{i+1,j} \end{aligned}$$

is stable, so that the error goes to zero as (i,j) moves away from the initial boundary. The solution is completed by shifting $\tilde{x}$ back to x.

For some experimental inputs and boundary conditions the true and observed semistates, and the errors are given in Fig. 1 - Fig. 6. □

References

[1] R. P. Roesser, "A discrete state-space model for linear image processing", IEEE Trans. Automat. Control, vol AC-20, pp. 1-10, Feb. 1975.

[2] G. Basille and G. Marro, "Controlled and conditioned-invariant subspaces in linear system theory," J. Opt. Theory App., vol. 3, pp. 306-315, 1969.

[3] E. Fornasini and G. Marchesini, "Doubly indexed dynamical systems: State-space models and structural properties," Mathematical systems Theory, v 12, pp 59-72, 1978.

[4] G. Conte, A.Perdon, and T. Kaczorek, "A geometric approach to singular 2-D linear syatems," Proc. IFAC Workshop on System Structure and Control, pp 241-244, Prague, Czechoslovakia, Sept. 1989.

[5] A.Karamancıoğlu and F. L. Lewis, "A geometric approach to 2-D implicit systems", Proceed. 29th CDC, pp. 470-475, Sept. 1990.

[6] A. Karamancıoğlu, 2-D Implicit Linear Systems, Ph.D. dissertation, Department of Electrical Engineering, University of Texas at Arlington, May 1991.

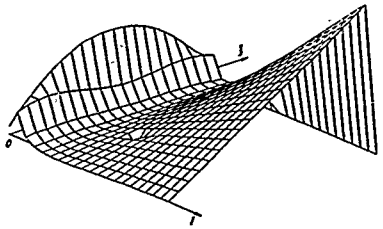


Fig. 1. True $5x^{\lambda 1} + 2x^{\lambda 2}$

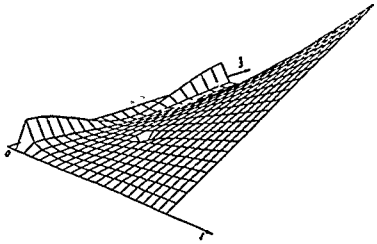


Fig. 2. Observed $5x^{\lambda 1} + 2x^{\lambda 2}$

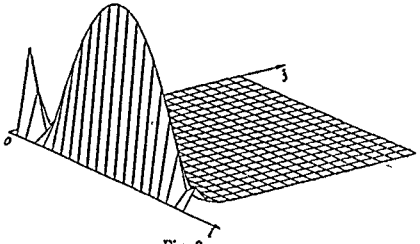


Fig. 3.
Error:= True $5x^{\lambda 1} + 2x^{\lambda 2}$ - Observed $5x^{\lambda 1} + 2x^{\lambda 2}$

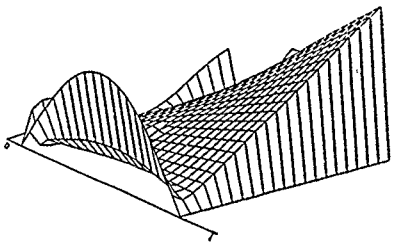


Fig. 4. True $4x^{\nu}$

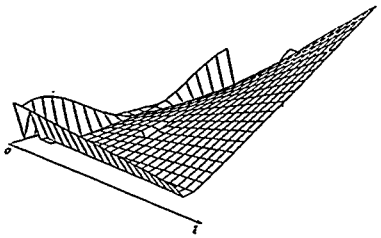


Fig. 5. Observed $4x^{\nu}$

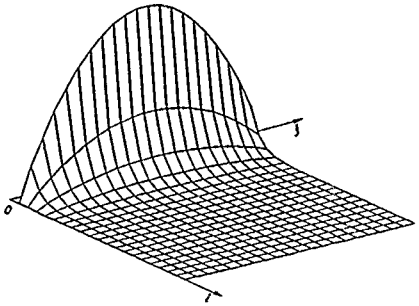


Fig. 6. Error:= True $4x^{\nu}$ - Observed $4x^{\nu}$

REGULARITY AND REGULARIZABILITY OF SINGULAR LINEAR SYSTEMS

J.E. Kurek
Instytut Automatyki Przemysłowej
Politechnika Warszawska
ul. Chodkiewicza 8, 02-525 Warszawa, Poland

Abstract. New conditions for regularity of singular linear system are presented. Then, regularizability problem for the system is considered. New conditions for system regularizability are given.

I. INTRODUCTION

Linear discrete-time time-invariant system can be described by the following equations:

$$\begin{aligned} Ex(k+1) &= Ax(k)+Bu(k) \\ y(k) &= Cx(k)+Du(k) \end{aligned} \qquad (1)$$

where $x \in R^n$, $u \in R^m$, $y \in R^p$ and E, A, B, C and D are real matrices of appropriate dimensions. If $E \neq I$ the system is named implicit or descriptor, [6], and in the consequence x is defined as a descriptor vector, u is an input vector and y is an output vector. If E is nonsquare or det A=0 the system is called singular.

Implicit and singular systems were considered in many papers, e.g. [1,3,4]. It follows from these papers that one of the most important and fundamental system properties is regularity.

In this note system regularity and regularizability are considered. New conditions for the existence of these properties are presented.

II. THE MAIN RESULT

A matrix pencil (A-sI) which is square and does not vanish identically is termed regular or nonsingular [2,9]. Consequently, the following definition was introduced.

Definition 1.
System (1) is regular iff (i) matrices E and A are square, and (ii) there exists s_o such that $\det(Es_o-A) \neq 0$.

□

The main feature of a regular singular system (1) is that the system is solvable and conditionable [7]. Moreover, almost all the known results related to singular system dynamics depend explicitly on the assumption of system regularity [5].

Now, the following theorem can be formulated.

Theorem 1.
System (1) is regular iff (i) dim E=n*n, and (ii)

$$\operatorname{rank}\begin{bmatrix} -A & E & 0 \\ 0 & -A & E \end{bmatrix} = 2n \qquad (2)$$

Proof.
Since system (1) is regular iff is solvable the condition for system regularity is that matrix F(N) has full rank for all N, [7], where

$$F(N) = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_N \end{bmatrix} = \begin{bmatrix} -A & E & 0 & 0 & \dots & 0 & 0 \\ 0 & -A & E & 0 & & & \vdots \\ \vdots & & \ddots & & \ddots & & \\ 0 & & \dots & & 0 & -A & E \end{bmatrix} \in R^{(N*n)*((N+1)*n)}$$

The matrix F(N) has less rows than columns. Therefore, it doesn't have a full rank iff its rows are linearly dependent. From the structure of matrix F(N) it follows that its rows are linearly independent iff there are linearly independent:
(i) rows of F_i,
(ii) rows of F_i and F_{i+1}, and
(iii) rows of F_i and F_{i-1}.
This is, however, equivalent to the given condition (ii).

Condition (i) is simply rewritten from the definition and has not to be proven.

□

Unfortunately, assumption of regularity is not preserved under linear feedback. Indeed, assuming system (1) is regular, in most cases it is quite easy to construct feedback u=Kx+v such that the closed loop singular system is nonregular. It was the motivation for introducing the regularizability notion in [8]. Here this notion is adapted for discrete-time system:

Definition 2.
System (1) is regularizable iff (i) matrices E and A are square, and (ii) there exist $K \in R^{m*n}$ and s_o such that $\det(Es_o-(A+BK)) \neq 0$.

□

The sense of the above is that there exists proportional feedback such that the closed loop singular system is regular. It is clear that this property is invariant with respect to feedback.

Then, the following theorem can be proven.

Theorem 2.
System (1) is regularizable iff (i) $E \in R^{n*n}$ and (ii)

$$\operatorname{rank}\begin{bmatrix} -A & E & 0 & B & 0 \\ 0 & -A & E & 0 & B \end{bmatrix} = 2n$$

Proof
Condition (i) is evident, it follows from the definition.

It is easy to see that always:

$$\operatorname{rank}\begin{bmatrix} -A-BK & E & 0 \\ 0 & -A-BK & E \end{bmatrix} \le \operatorname{rank}\begin{bmatrix} -A-BK & E & 0 & B & 0 \\ 0 & -A-BK & E & 0 & B \end{bmatrix} - \operatorname{rank}\begin{bmatrix} -A & E & 0 & B & 0 \\ 0 & -A & E & 0 & B \end{bmatrix}$$

Hence, the theorem 1 implies necessity of condition (ii). On the other hand, it is also clear that always exists $K_o \in R^{m*n}$ such that

$$\operatorname{rank}\begin{bmatrix} -A-BK_o & E & 0 \\ 0 & -A-BK_o & E \end{bmatrix} = \operatorname{rank}\begin{bmatrix} -A-BK_o & E & 0 & B & 0 \\ 0 & -A-BK_o & E & 0 & B \end{bmatrix} - \operatorname{rank}\begin{bmatrix} -A & E & 0 & B & 0 \\ 0 & -A & E & 0 & B \end{bmatrix}$$

This completes the proof of sufficiency of condition (ii).

□

Remark.
From the proof of theorem it is rather clear that condition (ii) is necessary and sufficient for the existence of matrices $F,K \in R^{m*n}$ such that $\det[(E-BF)s-(A+BK)] \neq 0$.

□

III. CONCLUDING REMARKS

New conditions for regularity and regularizability of singular system has been derived. The conditions are very easy and simple for use, based on theorem 1 the regularity of singular system can be verified simpler than using shuffle algorithm proposed by Luenberger in [7].

Finally, it should be emphasized that the conditions apply for discrete- as well for continuous-time system. It follows from remark to theorem 2 and fact that conditions for system regularity are the same for continuous- and discrete-time systems, [5].

ACKNOWLEDGMENT

This research was supported by Polish Ministry of National Education under Grant T/02/094/90-2

REFERENCES

[1] A.Banaszuk, M.Kociecki and K.M.Przyluski, "Remarks on duality between observation and control for implicit linear discrete time systems", Prepr. IFAC Workshop on System Structure and Control, Prague, Czechoslovakia, pp. 257-260, 1989.

[2] F.R.Gantmacher, The Theory of Matrices, vol. II, Chelsea, New York, 1953.

[3] T.Kaczorek, "The linear quadratic optimal regulator for singular 2-D system with variable coefficients", IEEE Trans. Autom. Control, vol. 34, pp. 565-566, 1989.

[4] J.E.Kurek, "On singular linear systems", IMACS Int. Symp. on Mathematical and Intelligent Models in System Simulation, pp. VI.A.4.1-4.4, Brussels, Belgium, 1990.

[5] F.L.Lewis, "A survey of linear singular systems", Circuits Sys. Sig. Process, vol. 5, pp. 3-36, 1986.

[6] D.G.Luenberger, "Dynamics equations in descriptor form", IEEE Trans. Automat. Control, vol. 22, pp. 312-321, 1977.

[7] D.G.Luenberger, "Time-invariant descriptor system", Automatica, vol. 14, pp. 473-480, 1978.

[8] K.Ozcaldiran and F.L.Lewis, "On the regularizability of singular systems", IEEE Trans. Automat. control, vol. 35, pp. 1156-1160, 1990.

[9] H.W.Turnbull and A.C.Aithen, An Introduction to theory of Canonical Matrices, Dover, New York, 1961.

INVERTIBILITY OF SINGULAR 2-D LINEAR SYSTEMS

Tadeusz KACZOREK
Accademia Polacca delle Scienze
2,Vicolo Doria,00187 Roma,Italy

Abstract-Two approaches for finding a whole class of inverse systems for a given singular 2-D linear system are presented.The first approach can be applied to regular 2-D linear systems with transfer matrices of full row ranks.The second approach can be also applied to nonregular singular 2-D linear systems.

I.INTRODUCTION

The problem of inverting linear and nonlinear multivariable systems has been under investigation for a long time [1,6,7,9,10,12].The first complete solution of the problem for regular linear systems was given by Silverman [11].Moylan[7]has considered the question of stability of the inverse system and he has modified the structure algorithm by introducing output and state-space transformations.The inversion of singular 1-D linear systems has been considered in[1,2,5,12].Lewis[2]and Beauchamp [1]have used the singular system structure algorithm to construct an inverse system for 1-D singular systems.
The main purpose of this paper is to present two approach for finding a whole class of inverse systems for a given singular 2-D linear system.

II.DEFINITION OF INVERSE SYSTEM

Consider the general singular model of 2-D linear systems[4]

$$\begin{aligned} Ex_{i+1,j+1} &= A_0x_{ij}+A_1x_{i+1,j}+A_2x_{i,j+1}+B_0u_{ij}+B_1u_{i+1,j}+B_2u_{i,j+1} \\ y_{ij} &= C_0x_{ij}+C_1x_{i+1,j}+C_2x_{i,j+1}+D_0u_{ij}+D_1u_{i+1,j}+D_2u_{i,j+1} \end{aligned} \quad (1)$$

where i,j are integer-valued horizontal and vertical coordinates,respectively,$x_{ij}\in R^n$ is the local semistate vector at the point(i,j),$u_{ij}\in R^m$ is the input,$y_{ij}\in R^p$ is the output and $E,A_k\in R^{nxn}$,$B_k\in R^{nxm}$,$C_k\in R^{pxn}$,$D_k\in R^{pxm}$,k=o,1,2 are real matrices.It is assumed that E is singular,i.e. det E =0 when q=n.
Let us consider an onther singular 2-D linear system of the form

$$\begin{aligned} E_{i+1,j+1} &= F_0x'_{ij}+F_1x'_{i+1,j}+F_2x'_{i,j+1}+G_0y_{ij}+G_1y_{i+1,j}+G_2y_{i,j+1} \\ u_{ij} &= H_0x'_{ij}+H_1x'_{i+1,j}+H_2x'_{i,j+1}+J_0y_{ij}+J_1y_{i+1,j}+J_2y_{i,j+1} \end{aligned} \quad (2)$$

where $x'_{ij}\in R^n$ is the local semistate vector,u_{ij} and y_{ij} are the same as in(1)and $E,F_k\in R^{nxn}$,$H_k\in R^{mxn}$,$J_k\in R^{mxp}$,k=o,1,2 are real matrices.
The system(2)is called an inverse system for(1)if the output y_{ij} of(1)is the input for(2)and the input u_{ij} of (1)is the output for(2).

III.REGULAR SINGULAR 2-D LINEAR SYSTEMS

The system(1)is called regular if and only if

$$\det[Ez_1z_2-A_0-A_1z_1-A_2z_2]\neq 0 \text{ for some } z_1,z_2\in C \quad (3)$$

where C is the field of complex numbers.
If(3)holds,then the transfer matrix of(1)is given by

$$T=T(z_1,z_2)=(C_0+C_1z_1+C_2z_2)[Ez_1z_2-A_0-A_1z_1-A_2z_2]^{-1}(B_0+B_1z_1+B_2z_2)+D_0+D_1z_1+D_2z_2 \quad (4)$$

If

$$\operatorname{rank} T(z_1,z_2)=p \text{ for some } z_1,z_2\in C \quad (5)$$

then there exists a right inverse $T_R=T_R(z_1,z_2)$of(4) given by[3]

$$T_R=T^t[TT^t]^{-1}+(I_m-T^t[TT^t]^{-1}T)Z \quad (6)$$

where the upper index t denotes the transposition,I_m is the mxm identity matrix and Z is an arbitrary mxp matrix Solving the equation $Y(z_1,z_2)=TU(z_1,z_2)$with respect to $U(z_1,z_2)$we obtain $U(z_1,z_2)=T_RY(z_1,z_2)$,where $U(z_1,z_2)$, $Y(z_1,z_2)$are the 2-D Z transforms of u_{ij} and y_{ij},respectively.Using one of the realization theory methods[3] for(6)we may find matrices E,F_k,G_k,H_k,J_k,k=o,1,2 of(2).

Note that the matrices depend on Z which is arbitrary. Therefore,we have shown the following

Theorem 1.

If the system(1)is regular and its transfer matrix(4) has full row rank,the there exists a whole class of inverse systems which can be found by the use of the realization theory methods.

Example 1.

Find an inverse system to(1)with

$$E=\begin{bmatrix}0&1&0\\0&0&0\\0&0&0\end{bmatrix},A_0=\begin{bmatrix}0&0&0\\0&0&-1\\0&0&0\end{bmatrix},A_1=\begin{bmatrix}-1&0&0\\0&1&0\\0&0&0\end{bmatrix},A_2=\begin{bmatrix}0&0&1\\0&0&0\\0&0&1\end{bmatrix},B_0=\begin{bmatrix}0&0\\0&0\\0&-1\end{bmatrix},$$

$$B_2=\begin{bmatrix}0&0\\-1&0\\0&0\end{bmatrix},B_1=0,C_0=[0\ 1\ 0],C_1=C_2=0,D_k=0 \text{ for } k=0,1,2$$

It is easy to check that the system is regular and

$$[Ez_1z_2-A_0-A_1z_1-A_2z_2]^{-1}=\frac{1}{z_1^2z_2}\begin{bmatrix}z_1z_2&z_1z_2^2&0\\0&-z_1z_2^2&-z_1z_2\\0&0&-z_1^2\end{bmatrix}$$

The transfer matrix(4)of the system is

$$T=\frac{1}{z_1^2z_2}[z_1z_2^2\ \ z_1]$$

and its right inverse is given by

$$T_R=\frac{1}{1+z_2^2}\begin{bmatrix}z_1z_2^3+p_1-p_2z_2^2\\z_1z_2-p_1z_2^2+p_2z_2^3\end{bmatrix} \quad (7)$$

where p_1 and p_2 are arbitrary.
Using the method given in[3]we obtain a realization of (7)in the form

$$E=\begin{bmatrix}0&0&1&0\\0&0&0&0\\0&0&0&0\\1&0&0&0\end{bmatrix},F_0=\begin{bmatrix}0&0&0&0\\0&0&1&0\\0&0&0&0\\0&1&0&0\end{bmatrix},F_1=\begin{bmatrix}-1&0&0&0\\0&0&0&0\\0&0&0&-1\\0&0&0&0\end{bmatrix},F_2=\begin{bmatrix}0&0&0&0\\0&0&0&-1\\0&1&0&0\\0&0&0&0\end{bmatrix},$$

$$G_0^t=[1\ 0\ 0\ 0],H_0=\begin{bmatrix}p_1&0&0&-p_2\\0&1&0&-p_1\end{bmatrix},H_1=\begin{bmatrix}0&0&1&0\\0&0&0&0\end{bmatrix},H_2=\begin{bmatrix}0&0&0&0\\0&0&p_2&0\end{bmatrix},$$

$G_0=G_2=0$, $J_k=0$ for k=o,1,2.

IV.GENERAL APPROACH

Using the matrices
$A:=[A_0,A_1,A_2],B:=[B_0,B_1,B_2],C:=[C_0,C_1,C_2],D:=[D_0,D_1,D_2]$
and the vectors

$$x_{ij}^t:=[x_{ij}^t,x_{i+1,j}^t,x_{i,j+1}^t],u_{ij}^t:=[u_{ij}^t,u_{i+1,j}^t,u_{i,j+1}^t]$$

we may write(1)in the form

$$Ex_{i+1,j+1}=Ax_{ij}+Bu_{ij} \quad (8a)$$

$$y_{ij}=Cx_{ij}+Du_{ij} \quad (8b)$$

Let rank E =r<q.Then there exists a nonsingular matrix $M\in R^{qxq}$ such that

$$ME=\begin{bmatrix}E_1\\0\end{bmatrix} \quad (9)$$

where $E_1\in R^{rxn}$ has full row rank.
Premultiplying(8a)by M and using(9)we obtain

$$E_1x_{i+1,j+1}=\bar{A}_1x_{ij}+\bar{B}_1u_{ij} \quad (10a)$$

$$0=\bar{A}_2x_{ij}+\bar{B}_2u_{ij} \quad (10b)$$

where

$$\begin{bmatrix}\bar{A}_1\\\bar{A}_2\end{bmatrix}:=MA,\ \bar{A}_1\in R^{rx3n},\ \bar{A}_2\in R^{(q-r)x3n},\ \begin{bmatrix}\bar{B}_1\\\bar{B}_2\end{bmatrix}:=MB,\ \bar{B}_1\in R^{rx3m},\ \bar{B}_2\in R^{(q-r)x3m}$$

The equations(8b)and(10b)may be written as

$$\begin{bmatrix}y_{ij}\\0\end{bmatrix}=\begin{bmatrix}C\\\bar{A}_2\end{bmatrix}x_{ij}+\begin{bmatrix}D\\\bar{B}_2\end{bmatrix}u_{ij} \quad (11)$$

Let

$$\operatorname{rank}\begin{bmatrix}D\\\bar{B}_2\end{bmatrix}=r_1<p+q-r$$

Then there exists a nonsingular matrix

$$S=\begin{bmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{bmatrix}\in R^{(p+q-r)x(p+q-r)} \qquad \begin{matrix}S_{11}\in R^{r_1xp}\\ S_{21}\in R^{(p+q-r-r_1)xp}\end{matrix}$$

such that

$$S\begin{bmatrix}D\\ B_0\end{bmatrix}=\begin{bmatrix}\bar{D}\\ 0\end{bmatrix} \tag{12}$$

where $\bar{D}\in R^{r_1x3n}$ has full row rank.

Premultiplying (11) by S and using (12) we obtain

$$S_{11}y_{ij}=\bar{C}_1x_{ij}+\bar{D}u_{ij} \tag{13a}$$

$$S_{21}y_{ij}=\bar{C}_2x_{ij} \tag{13b}$$

where

$$\begin{bmatrix}\bar{C}_1\\ \bar{C}_2\end{bmatrix}:=-S\begin{bmatrix}C\\ A_2\end{bmatrix}, \qquad \begin{matrix}\bar{C}_1\in R^{r_1x3n}\\ \bar{C}_2\in R^{(p+q-r-r_1)x3n}\end{matrix}$$

A right inverse D_R of $\bar{D}$ is given by [8]

$$D_R=\bar{B}^t[\bar{B}\bar{B}^t]^{-1}+(I_{3n}-\bar{B}^t[\bar{B}\bar{B}^t]^{-1}\bar{B})P \tag{14}$$

where P is an arbitrary $3mxr_1$ matrix.

From (13a) we have

$$u_{ij}=-D_R\bar{C}_1x_{ij}+D_RS_{11}y_{ij} \tag{15}$$

Note that (10a) and (13b) can be written as

$$\begin{bmatrix}E_1\\ 0\end{bmatrix}x_{i+1,j+1}=\begin{bmatrix}A_1\\ \bar{C}_2\end{bmatrix}x_{ij}+\begin{bmatrix}B_1\\ 0\end{bmatrix}u_{ij}+\begin{bmatrix}0\\ -S_{21}\end{bmatrix}y_{ij} \tag{16}$$

Substitution of (15) into (16) yields

$$\begin{bmatrix}E_1\\ 0\end{bmatrix}x_{i+1,j+1}=\begin{bmatrix}A_1-B_1D_R\bar{C}_1\\ \bar{C}_2\end{bmatrix}x_{ij}+\begin{bmatrix}B_1D_RS_{11}\\ -S_{21}\end{bmatrix}y_{ij} \tag{17}$$

From comparison of (17) and (2) we have

$$E=\begin{bmatrix}E_1\\ 0\end{bmatrix},F:=[F_o,F_1,F_2]=\begin{bmatrix}A_1-B_1D_R\bar{C}_1\\ \bar{C}_2\end{bmatrix},G_o:=\begin{bmatrix}B_1D_RS_{11}\\ -S_{21}\end{bmatrix}, \tag{18}$$

$G_1=G_2=0$, $q:=p+q-r_1$.

Let

$$D_R=[D_{R1}^t,D_{R2}^t,D_{R3}^t]^t, \quad D_{Rk}\in R^{mxr_1} \text{ for } k=0,1,2$$

Then from (15) we get

$$u_{ij}=-D_{R1}\bar{C}_1x_{ij}+D_{R1}S_{11}y_{ij} \tag{19}$$

and from comparison of (19) and (2) we have

$$H:=[H_o,H_1,H_2]=-D_{R1}\bar{C}_1, \quad J_o:=D_{R1}S_{11}, J_1=J_2=0 \tag{20}$$

Therefore we have proved the following

Theorem 2.

If $r_1>0$, then there exists for (1) an inverse system (2) with matrices given by (18) and (20).

If $r_1=0$, then matrices of (2) can be found by the use of the following

Procedure:

Step 1. Find M satisfying (9) and $E_1,\bar{A}_1,\bar{A}_2,\bar{B}_1,\bar{B}_2$.

Step 2. Find S satisfying (12) and $\bar{D},\bar{C}_1,\bar{C}_2,D_R$.

Step 3. Using (18) and (20) find F_k,G_k,H_k and J_k for k=o,1,2.

Example 2.

Find an inverse system to (1) with

$$E=\begin{bmatrix}0&1&0\\0&0&0\\0&0&1\end{bmatrix},A_o=0,A_1=\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix},A_2=\begin{bmatrix}0&0&0\\0&0&1\\0&0&0\end{bmatrix},B_o=\begin{bmatrix}0&1\\1&0\\0&0\end{bmatrix},B_1=B_2=0, \tag{21}$$

$$C=\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix},C_1=C_2=0,\ D_o=\begin{bmatrix}0&0\\0&1\end{bmatrix},\ D_1=D_2=0$$

It is easy to check that the condition (3) is not satisfied for (21). Therefore, the first approach can not be applied.

Using the Procedure we obtain:

Step 1. The matrix M satisfying (9) has the form

$$M=\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix} \text{ and } E_1=\begin{bmatrix}0&1&0\\0&0&1\end{bmatrix},$$

$$\begin{bmatrix}\bar{A}_1\\ \bar{A}_2\end{bmatrix}=MA=\begin{bmatrix}0&0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&1\end{bmatrix}\begin{bmatrix}\bar{B}_1\\ \bar{B}_2\end{bmatrix}=M\bar{B}=\begin{bmatrix}0&1&0&0&0&0\\0&0&0&0&0&0\\1&0&0&0&0&0\end{bmatrix} \tag{22}$$

Step 2. The matrix S satisfying (12) has the form

$$S=\begin{bmatrix}S_{11}&S_{12}\\ S_{21}&S_{22}\end{bmatrix}=\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix} \text{ and } \bar{D}=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\end{bmatrix},$$

$$\begin{bmatrix}\bar{C}_1\\ \bar{C}_2\end{bmatrix}=S\begin{bmatrix}C\\ A_2\end{bmatrix}=\begin{bmatrix}0&0&0&0&0&0&0&0&1\\0&1&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0\end{bmatrix} \tag{23}$$

Taking into account that $\bar{D}\bar{D}^t=I_2$ from (14) we obtain

$$D_R=\begin{bmatrix}I_2\\ P_2\end{bmatrix}, P_2 \text{ is arbitrary } 4x2 \text{ matrix} \tag{24}$$

Step 3. Using (18), (20), (22), (23) and (24) we obtain

$$E=\begin{bmatrix}E_1\\ 0\end{bmatrix}=\begin{bmatrix}0&1&0\\0&0&1\\0&0&0\end{bmatrix},F=[F_o,F_1,F_2]=\begin{bmatrix}\bar{A}_1-\bar{B}_1D_R\bar{C}_1\\ \bar{C}_2\end{bmatrix}=$$

$$=\left[\begin{array}{ccc|ccc|ccc}0&-1&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&-1\end{array}\right],G_o=\begin{bmatrix}\bar{B}_1D_RS_{11}\\ -S_{21}\end{bmatrix}=\begin{bmatrix}0&1\\0&0\\-1&0\end{bmatrix},G_1=G_2=0,$$

$$H=[H_o,H_1,H_2]=-D_{R1}\bar{C}_1=\left[\begin{array}{ccc|ccc|ccc}0&0&0&0&0&0&0&-1&0\\0&-1&0&0&0&0&0&0&0\end{array}\right],J_o=D_{R1}S_{11}=\begin{bmatrix}0&0\\0&1\end{bmatrix}$$

$J_1=J_2=0$

Hence the desired inverse system has the form

$$\begin{bmatrix}0&1&0\\0&0&1\\0&0&0\end{bmatrix}x_{i+1,j+1}=\begin{bmatrix}0&-1&0\\0&0&0\\1&0&0\end{bmatrix}x_{ij}+\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}x_{i+1,j}+\begin{bmatrix}0&1\\0&0\\-1&0\end{bmatrix}y_{ij}$$

$$u_{ij}=\begin{bmatrix}0&0&0\\0&-1&0\end{bmatrix}x_{ij}+\begin{bmatrix}0&0&-1\\0&0&0\end{bmatrix}x_{i,j+1}+\begin{bmatrix}0&0\\0&1\end{bmatrix}y_{ij}$$

With slight modifications both approaches can be extended for n=D/n>2/ singular linear systems.

REFERENCES

[1] G.Beauchamp, Algorithms for singular systems, Ph.D. thesis at Georgia Inst. Techn., January 1990.

[2] M.El-Tohami, On minimal order inverses of discrete-time descriptor systems. Int.J.Contr. vol.41, No.4, 1985, 991-104.

[3] T.Kaczorek, Realization problem for singular 2-D linear discrete systems, Bull.Pol.Acad.Techn.Sci. vol.37, No.1-2, 1989, 37-48.

[4] T.Kaczorek, The singular general model of 2-D systems and its solution, IEEE Trans.Autom.Contr. AC-33, Nov.1988, 1060-1061.

[5] F.Lewis, Inversion of descriptor systems. Proc.American Control.Conf., June 1983, 1153-1158.

[6] J.L.Massey, M.K.Sain, Inversion of linear sequential systems, IEEE Trans.Comp. C-17, April 1968, 330-337.

[7] P.J.Moylan, Stable inversion of linear systems, IEEE Trans.Autom.Contr. AC-22, Feb.1974, 74-78.

[8] C.R.Rao, S.K.Mitra, Generalized Inverse of Matrices and Its Applications, New York, Wiley, 1971.

[9] W.Respondek, Right and left invertibility of nonlinear control systems, Proc.Workshop Control.and Opt.Contr. Rutgers Univ., May 1987, Ed.R.Sussman, Marcel Dekker.

[10] M.K.Sain, J.L.Massey, Invertibility of linear time-invariant dyramical systems, IEEE Trans.Autom.Contr. AC-14, April 1969, 141-149.

[11] L.M.Silverman, Inversion of multivariable linear syst-ems, IEEE Trans.Autom.Contr. AC-14, June 1969, 270-276.

[12] S.Tan, J.Vandewalle, Inversion of singular systems, IEEE Trans.Autom.Contr. AC-33, May 1988, 583-587.

COMPLETE CONTROLLABILITY OF SINGULAR 2-D SYSTEMS

JERZY KLAMKA

Institute of Automation, Technical University
44-100 Gliwice , Poland

Abstract - The general singular 2-D linear systems with variable coefficients are considered. The concept of complete controllability is extended for the singular model and variable coefficients. Using the general response formula necessary and sufficient condition for complete controllability in a given rectangular is established.

I. SYSTEM DESCRIPTION

Let us consider the general model of singular linear 2-D system with variable coefficients given by the following equation [4]:

$$E(i+1,j+1)x(i+1,j+1) = A_0(i,j)x(i,j) + A_1(i+1,j)x(i+1,j) + A_2(i,j+1)x(i,j+1) + B(i,j)u(i,j) \qquad /1/$$

where i,j are integer-valued vertical and horizontal coordinates, respectively, $x(i,j) \in R^n$ is the local semistate vector at (i,j), $u(i,j) \in R^m$ is the input vector, and $A_k(i,j)$, $k=1,2,0$, $B(i,j)$ are real matrices of appropriate dimensions with entries depending on i and j. The special feature of the model /1/ is that the matrix $E(i,j)$ may be singular, [2], [3], [4].

Boundary conditions for /1/ are given by [4]:

$$x(i,j_0) = x_{ij_0} \quad , \quad i=i_0,i_0+1,i_0+2,\ldots$$
$$x(i_0,j) = x_{i_0j} \quad , \quad j=j_0,j_0+1,j_0+2,\ldots \qquad /2/$$

for $i_0, j_0 = 0,1,2,\ldots$, where x_{ij_0} and x_{i_0j} are known vectors.

It is assumed that the system /1/ is solvable and boundary conditions are admissible [2], [4]. The transition matrix G_{pq}^{ij} for system /1/ is defined as follows [4]:

$$E(i+1,j+1)\,G_{i-p+1,j-q+1}^{i+1,j+1} = A_0(p,q)\,G_{00}^{pq} + A_1(p+1,q)G_{1,0}^{p+1,q} + A_2(p,q+1)G_{0,1}^{p,q+1} + I \quad \text{for } i=p,\ j=q$$

$$E(i+1,j+1)\,G_{i-p+1,j-q+1}^{i+1,j+1} = 0 \qquad /3/$$

for $p \le i+n_1+1$ and $q \le j+n_2+1$, $n_1 \le n$, $n_2 \le n$

$$E(i+1,j+1)\,G_{i-p+1,j-q+1}^{i+1,j+1} = A_0(i,j)G_{i-p,j-q}^{i,j} + A_1(i+1,j)G_{i-p+1,j-q}^{i+1,j} + A_2(i,j+1)G_{i-p,j-q+1}^{i,j+1}$$

for $i \neq p$ and/or $j \neq q$.

Using the transition matrix G_{pq}^{ij} the exact formula for the solution of the system /1/ with admissible boundary conditions /2/ can be derived /see the paper [4] for details/.

II. BASIC DEFINITIONS

Now we shall introduce the concept of complete controllability of the system /1/.

Definition. The system /1/ is said to be completely controllable in the rectangular $[(i_0,j_0),(r,s)]$ if for any admissible boundary conditions /2/ and every vectors $x_{rl} \in R^n$, $l=j_0+1,j_0+2,\ldots,s-1$, $x_{ks} \in R^n$, $k=i_0+1,i_0+2,\ldots,r-1$, $x_{rs} \in R^n$, there exists a sequence of input vectors $u(i,j)$ for $(i_0,j_0) \le (i,j) \le (r+n_1 , s+n_2)$ such that

$$x(r,l) = x_{rl} \quad \text{for } l=j_0+1,j_0+2,\ldots,s-1$$
$$x(k,s) = x_{ks} \quad \text{for } k=i_0+1,i_0+2,\ldots,r-1 \qquad /4/$$
$$x(r,s) = x_{rs}$$

It is easily to verify, that complete controllability in a given rectangular implies local controllability in the same rectangular. The converse statement is not always true [1] [5]

In order to formulate the necessary and sufficient condition for complete controllability let us introduce the following notations:

$$Q_{rj} = \left[G_{r-i_0,j-j_0}^{r,j}B(i_0,j_0) \,\middle|\, G_{r-i_0-1,j-j_0}^{r,j}B(i_0+1,j_0) \,\middle|\, \ldots \,\middle|\, G_{1,j-j_0}^{r,j}B(r-1,j_0) \,\middle|\, G_{0,j-j_0}^{r,j}B(r,j_0) \,\middle|\, \ldots \,\middle|\, G_{-n_1,j-j_0}^{r,j}B(r+n_1,j_0) \right] \qquad /5/$$

Q_{rj} are $n \times m(r+n_1-i_0)$-dimensional matrices for $j = j_0,j_0+1,\ldots,s+n_2-1$.

In a quite similar way we may define the matrices $\tilde{Q}_{pq}$, for $p = i_0+1,i_0+2,\ldots,r$ and $q=j_0,j_0+1,j_0+2,\ldots,s-1$.

$$\tilde{Q}_{pq} = \left[G^{i,j}_{i-p,j+q+1-s}B(p,s-q-1) \;\middle|\; G^{i,j}_{i-p+1,j+q+1-s}B(p-1,s-q-1) \;\middle|\; \ldots \;\middle|\; G^{i,j}_{i-i_0,j+q+1-s}B(i_0,s-q-1) \;\middle|\; 0 \;\middle|\; \ldots \;\middle|\; 0 \right] \quad /6/$$

III. CONTROLLABILITY CONDITION.

Using matrices Q_{ri} and Q_{pq} we may define the so called complete controllability matrix $W^{r,s}_{i_0,j_0}$ for the system /1/ and given rectangular $[(i_0,j_0),(r,s)]$.

$$W^{r,s}_{i_0,j_0} = \begin{bmatrix} Q_{r,j_0} & 0 & \cdots \\ Q_{r,j_0} & Q_{r,j_0+1} & \cdots \\ \cdots & \cdots & \cdots \\ Q_{r,j_0} & Q_{r,j_0+1} & \cdots \\ \tilde{Q}_{i_0+1,j_0} & \tilde{Q}_{i_0+1,j_0+1} & \cdots \\ \tilde{Q}_{i_0+2,j_0} & \tilde{Q}_{i_0+2,j_0+1} & \cdots \\ \cdots & \cdots & \cdots \\ \tilde{Q}_{r,j_0} & \tilde{Q}_{r,j_0+1} & \cdots \end{bmatrix}$$

$$\begin{bmatrix} \cdots & 0 & 0 \\ \cdots & 0 & 0 \\ \cdots & \cdots & \cdots \\ \cdots & Q_{r,s+n_2-2} & Q_{r,s+n_2-1} \\ & Q_{i_0+1,s+n_2-2} & Q_{i_0+1,s+n_2-1} \\ \cdots & \tilde{Q}_{i_0+2,s+n_2-2} & \tilde{Q}_{i_0+1,s+n_2-1} \\ \cdots & \cdots & \cdots \\ \cdots & \tilde{Q}_{r,s+n_2-2} & \tilde{Q}_{r,s+n_2-1} \end{bmatrix} \quad /7/$$

$W^{r,s}_{i_0,j_0}$ is $n(r+s-1-i_0-j_0) \times m(r+n_1-i_0) \cdot (s+n_2-j_0)$ -dimensional constant matrix.

Theorem. The dynamical system /1/ is completely controllable in the rectangular $[(i_0,j_0),(r,s)]$ if and only if

$$\operatorname{rank} W^{i_0,j_0}_{r,s} = n(r+s-1-i_0-j_0) \qquad /8/$$

Proof. We shall present only the sketch of the proof. Using the general response formula we may obtain the relations between the vectors /4/ and the control sequence in the rectangular $[(i_0,j_0),(r+n_1, s+n_2)]$. This relation is given by the complete controllability matrix. Hence, the full row rank of the complete controllability matrix is the necessary and sufficient condition of the existence of control sequence, which steers the dynamical system /1/ to the desired final states given by the formulas /4/.

IV. CONCLUSIONS

In the paper using complete controllability matrix necessary and sufficient condition for complete controllability in a given rectangular of linear singular 2-D system is formulated. The above considerations can be also extended for the general model of M-D linear singular systems with variable coefficients. Using the complete controllability matrix it is possible to solve analytically the so called minimum energy control problem, [1], [3].

REFERENCES

[1] Kaczorek T., Two-Dimensional Linear Systems, Springer-Verlag, Berlin, 1985.

[2] Kaczorek T., Singular general model of 2-D systems and its solution, IEEE Transactions Autom. Control, vol. AC-33, no. 10, 1988, 1060-61.

[3] Kaczorek T., General response formula and minimum energy control for the general singular model of 2-D systems, IEEE Trans. Autom, Control, vol. AC-35, no. 4, 1990, 433-436.

[4] Kaczorek T., General response formula for singular 2-D linear systems with variable coefficients, Private communication, 1990.

AN IMAGE PROCESS FOR ACHIEVING WYSIWYG COLOUR

WAAD S. YOUSIF
Department of Computer Studies
Loughborough University of Technology
Loughborough
LEICS. UK

AND

M. RONNIER LUO
Department of Computer Studies
Loughborough University of Technology
Loughborough
LEICS. UK

Abstract The work described in this paper summarises the results obtained from a research project. A model of colour vision, Hunt-Alvey Colour Appearance Model, was derived to accurately predict colour appearance under a wide range viewing conditions on both self-luminance display and reflection print media. The model gave good prediction to a comprehensive experimental database which was also obtained in this project. A method for characterisation imaging devices, i.e. self-luminance displays and electronics printers, was established and its associated mathematical models were also devised.

By integrating all the results, an image process, named the "four-stage transform", was formed and the processed image presented on a hardcopy gave a good appearance match to that displayed on a monitor. This implies that the dream of WYSIWYG colour could become a reality.

I. INTRODUCTION

The range of computer supported applications where colour selection and manipulation is a natural characteristic of the user's task is large, and includes such disciplines as graphics, architectural, textile, product, interior, medical, and engineering design. With the proliferation of colour display capability on computers and the recent availability of cheaper digital colour printers, the market is growing even faster. The typical problem of these systems is that what users see on one display cannot be faithfully reproduced onto another, and cannot be reasonably presented on a hardcopy. This is a serious problem. For example, a computer user may spend hours painstakingly choosing and balancing colours on screen in order to create the desired image, then takes only a minute to print out a hardcopy. Yet it is the hardcopy that is often the primary communication tool used for judgments about the design. The outcome from poor fidelity of hardcopy could be both practical and aesthetic. Hence "What You See Is What You Get", WYSIWYG, has been an elusive target for the computer colour industry.

There are two causes of this problem. First, there is a lack of understanding of the properties of human colour perception in the different viewing conditions used for displays and reflection prints. The second cause of this problem is that the colour primaries and colour generating technologies of displays and printers are very different. With all this in mind, a consortium was formed in 1986 to tackle above problems. This research project entitled "Predictive perceptual colour models" was carried out under the auspices of the UK government's Alvey programme. The member of consortium were Crosfield Electronics, Sigmex Displays and LUTCHI . (LUTCHI stands for Loughborough University of Technology Computer-Human Interface Research Centre.) The objectives of the project were to:

- Derive a computer-based model of colour vision to predict the colour appearance under various viewing conditions, i.e. illuminants, luminance level, background, and media (display and reflection print).
- Develop practical methods for characterising different colour printing and display devices, so that proofing simulation is possible.

II. DERIVING HUNT-ALVEY COLOUR APPEARANCE MODEL

The work for achieving the first goal was divided into three stages which were: 1) to conduct a large scale psychological experiment for assessing colour appearance under various viewing conditions by a panel of normal colour vision subjects. 2) to implement existing available colour models and to compare its performance using the data obtained from stage one, and 3) to modify a particular model which performed the best from stage two until the best fit was achieved to the data.

The first stage of work was the acquisition of experimental data. Ten observers were asked to make judgements of lightness, colourfulness, and hue of a sequence of test patches at the centre of a complex viewing field, using a magnitude estimation technique. Painted colour chips stuck onto a grey card were presented in a viewing cabinet and its replicate displayed on a high resolution colour display. The decorating patches around the periphery of the field were selected randomly to create a "mondrian" pattern to simulate a complex field within which to assess the test colour. The overall experiment was divided into 23 phases, with approximately 100 colours in each phase, to allow four different parameters to be studied:

- four types of illuminants (D65, D50, white fluorescent, tungsten)
- two luminance levels (high (250 cd/m^2) and low (40 cd/m^2))
- three backgrounds (white, grey, and black)
- two displayed modes (self-luminace display and reflection prints)

Altogether this generated an experimental database containing over 43,000 estimations, one of the largest such body of data to be amassed anywhere in the world since 1945. This data set is named the LUTCHI Colour Appearance Data [1].

At the second stage of work, this data was used to test the predictive accuracy of various existing colour models. For five models tested, the Hunt's model [2] outperformed the others and gave reasonably accurate prediction to the data. This model was further refined to become Hunt-Alvey Colour Appearance Model (Hunt-ACAM) [3]. The errors of prediction from this model are similar to those between each individual observer's and mean visual results. This implies that the model's performance is close to that of typical observing variation. A block diagram of Hunt-ACAM is given in Figure 1 to illustrate its functions. The input parameters are CIE XYZ tristimulus values [4], which are physical quantities to define a particular colour in question, together with the information of illuminant, luminance level, background, and display mode. (The tristimulus values are a set of values for specifying a colour and recommended to be used by the Commission Internationale d'Eclairage (CIE). The X, Y, and Z are the amount of red, green, and blue lights in order to match a colour in question.) The output information from the model are the predictive perceived attributes, i.e. lightness, brightness, colourfulness, chroma, saturation, and hue [5]. The reverse model has also been derived to transform the lightness (L), colourfulness(C), and hue (H) of a colour under a particular set of viewing conditions to the corresponding CIE XYZ values.

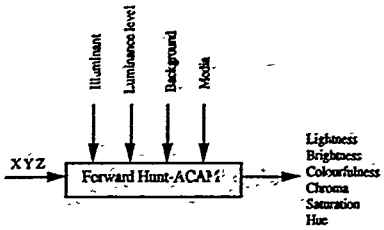


Figure 1. A block diagram to illustrate the Hunt-Alvey Colour Appearance Model (Hunt-ACAM)

III. DEVICES CHARACTERISATION

For tackling the second cause of poor fidelity problem, we developed a method for the characterisation of colour display and printing devices. We generate a "colour cube" with nine levels for each of the three colour primaries of the device, i.e. red, green and blue for the display, and cyan, magenta, and yellow for the printers. In total, 9x9x9 = 729 samples were produced for each device.

For a colour printer, or proofing system such as Cromalin, with cyan, magenta, and yellow inks, this method included printing nine charts each containing 9x9 samples, as shown in Figure 2. These samples were then measured with a Macbeth MS2000 spectrophotometer to obtain their absorption spectra, from which the CIE XYZ tristimulus values were calculated.

For a colour monitor, with red, green, and blue drive signals, a corresponding sequence of colour patches is generated at the centre of the screen via a telespectroradiometer (TSR) to obtain their emission spectra, from which again the XYZ tristimulus values were calculated.

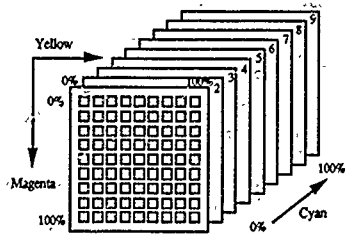


Figure 2. Arrangement of colour patches in device calibration cube (using cyan, magenta and yellow ink primaries)

The Foward and Reverse Device Models were developed in order to calculate the XYZ values for any intermediate values of CMY (or RGB) and vice versa. The formula is given in Mitchell and Wait[6] and used by Stone et al [7].

For the forward device model, the transformation of the device coordinates, say (r,g,b), into tristimulus values (X,Y,Z) was carried out by using the transformations

$$t = \sum_{i=1}^{8} \varphi_i(r,g,b)\, t_i, \qquad (t = X,Y,Z) \qquad (1)$$

where the functions $\varphi_i(r,g,b)$, (i=1,2,...,8), are given by

$\varphi_1 = (1-D_r)(1-D_g)(1-D_b)$, $\varphi_2 = D_r(1-D_g)(1-D_b)$, etc.,

where D_r, D_g and D_b are the distances from the origin of the selected cube to the desired value, normalised to lie in the range 0-1. The vertices are numbered as in Figure 3.

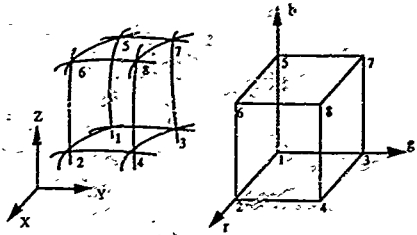


Figure 3

An arbitrary hexahedron is thus transformed into the unit cube in (r,g,b) space. The interpolation formula used to obtain the value of X corresponding to a colour produced by a set of (r,g,b) coordinates, where (r,g,b) is known to lie in this cube, is then defined by

$$X(r,g,b) = \sum_{i=1}^{8} \varphi_i(r,g,b)\, X_i$$
$$= (1-D_r)(1-D_g)(1-D_b)X_1 + D_r(1-D_g)(1-D_b)X_2 + (1-D_r)D_g(1-D_b)X_3 + D_rD_g(1-D_b)X_4 + (1-D_r)(1-D_g)D_bX_5 + D_r(1-D_g)D_bX_6 + (1-D_r)D_gD_bX_7 + D_rD_gD_bX_8 \qquad (2)$$

Similarly the Y and Z tristimulus values were calculated.

For the reverse device model for finding ,say (c,m,y) from given tristimulus values X,Y,Z, we start from the middle cube which has an origin of (5,5,5) and find the corresponding c,m,y and X,Y,Z values at that point. We then calculate ΔX, ΔY, ΔZ, and obtain the differences in c, m, y values by solving

$$(\Delta c\ \Delta m\ \Delta y)^T = J^{-1}\,(\Delta X\ \Delta Y\ \Delta Z)^T \qquad (3)$$

subsequently,

$$(c\ m\ y)^{(k+1)T} = (c\ m\ y)^{(k)T} + (\Delta c\ \Delta m\ \Delta y)^{(k)T} \qquad (4)$$

where k is the iteration parameter and J is the Jacobian matrix given by

$$J = \begin{bmatrix} \frac{dX}{dc} & \frac{dX}{dm} & \frac{dX}{dy} \\ \frac{dY}{dc} & \frac{dY}{dm} & \frac{dY}{dy} \\ \frac{dZ}{dc} & \frac{dZ}{dm} & \frac{dZ}{dy} \end{bmatrix}$$

By applying the interpolation formula, see equation(2), a new set of (X, Y, Z) values are then calculated and $(\Delta X\ \Delta Y\ \Delta Z)^{(k+1)}$ is found. If these values are within tolerance we exit with the calculated c, m and y values of equation(4). Otherwise, we proceed to the next iteration, equation(3) , using the same matrix J if still in the same cube. If it is not in the same cube we start the same procedure with the new calculated point (new cube), until convergence is achieved.

IV. IMAGE PROCESSING

The real achievement of the project has been the ablity to combine the device characterisation method with the colour appearance model into a suite of image processing software. This allows us not only to simulate the appearance of an image when produced on a different device but also to predict the change in appearance of a coloured image under various different conditions. When incorporated into a computer-based design system this technology would allow the users to visualise on the display monitor the true colour appearance of the final product and to have a high degree of confidence that this colour specification would be preserved into production.

This image processing software is designated the "Four-Stage Transform" (FST) and it is illustrated in Figure 4. For this particular example, the operator wishes to produce a "hardcopy" version of the displayed image. Initially, the image represented in the red, green, and blue primaries of a display was converted, pixel by pixel, into an equivalent XYZ image using the **Forward Device Model**. Secondly, this XYZ image was then transformed by the **Forward Hunt-ACAM** to obtain the LCH image in which the LCH are the perceived attributes under the display viewing conditions. Thirdly, the LCH image was converted to an XYZ image which preserves the same appearance but viewed under the hardcopy viewing conditions using **Reverse Hunt-ACAM**. Finally, the XYZ image was processed to achieve cyan, magenta, and yellow ink primaries of a printer using the **Reverse Device Model**.

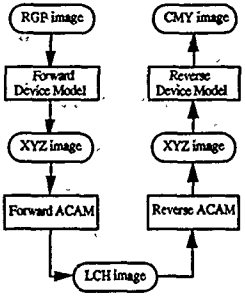


Figure 4. A block diagram to illustrate the four-stage transform

Several images have been tested. The results are quite satisfactory that good appearance match between the printed and displayed image was obtained.

V. WHAT'S IN STORE?

We commenced a further three-year research project in February 1990. The new consortium consists of Crosfield Electronics and Loughborough University, as before, but also includes Coats Viyella, the giant textile manufacturer. The aims of the new project are:

- to extend the functionality of Hunt-ACAM into new media and viewing conditions.
- to improve the device characterisation methods.
- to develop better user interfaces for the management of colours.
- to bring the image processing technology nearer to commercial exploitation.

We are confident that this on-going research effort will yield great benefits in the medium term for management of colour in computer-based design systems, and will enable the dream of WYSIWYG colour to become reality.

References

1. M.R.Luo, A.A.Clarke, P.A.Rhodes, A.Shappo, A.A R. Scrivener and C.Tait, Quantifying Colour Appearance - Part I LUTCHI Colour Appearance Data, Color Res. Appl. 16, 000-000, 1991.
2. R.W.G Hunt, "A model of colour vision for predicting colour appearance in various viewing conditions, Color Res. Appl. 12, 297-314, 1987.
3. M R.Luo, A.A.Clarke, P.A.Rhodes, A.Shappo, A.A.R. Scrivener, and C.Tait, Quantifying Colour Appearance - Part II Testing Colour Models Performance Using LUTCHI Colour Appearance Data, Color Res. Appl. 16, 000-000, 1991.
4. CIE, Colorimetry, second edition, CIE Publication No. 15.2, Central Bureau of the CIE, Paris, 1986.
5. CIE International Lighting Vocabulary, CIE Publication No. 17.4, Central Bureau of the CIE, Geneva, Switzerland, 1987.
6. A.R.Mitchell and R.Wait, The Finite Element Method in Partial Differential Equations, John Wiley & Sons, Chichester, 1977.
7. M.C.Stone, W.B.Cowan and J.C.Beatty, Color Gamut Mapping and the Printing of Digital Color Images, ACM Transaction Graphics. 7, 249-292, 1988.

HISTOGRAM SEGMENTATION FILTERING: A GRAPHICS DCT COMPRESSION POSTPROCESSOR

R.A.KIRK
Crosfield Electronics Limited
Three Cherry Trees Lane
Hemel Hempstead HP2 7RH
England.

Abstract:- Many DCT compression systems work well on scanned images but cannot compress text and other artificial features without injecting visible errors. These features generally have sharp edges between flat regions. The histogram of the pixel values near such a feature will typically show a few sharp peaks. If we segment the histogram of a DCT block into its peaks we can identify these features and reconstruct the original values. Crosfield DCT compression results are presented.

1. INTRODUCTION

The Discrete Cosine Transform (DCT) of a small (8*8) section of an image will often have many values close to zero. DCT compression techniques [1]-[4] exploit this feature by quantizing these values to give a compactly codeable set of integers, most of them zero. This introduces small errors. The quantization is usually optimised to make the errors invisible under normal conditions.

Sharp edges between flat regions present particular problems for a DCT compression system. A sharp feature will have many large values in its transform, so the quantization errors may be unusually severe. The errors will be unusually visible against the flat background.

Fig. 1. Experimental DCT compression of artificial test image without Histogram Segmentation Filtering

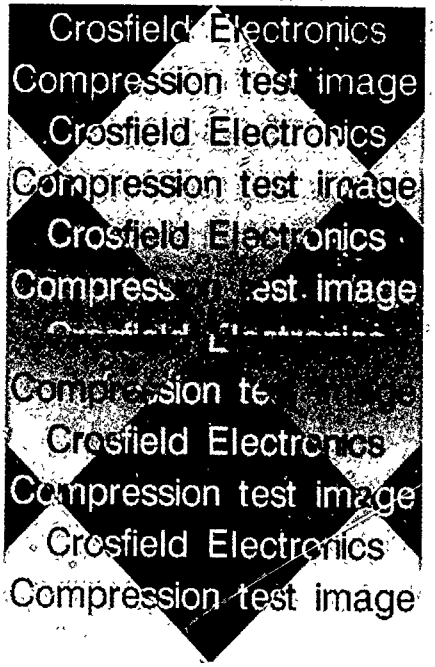
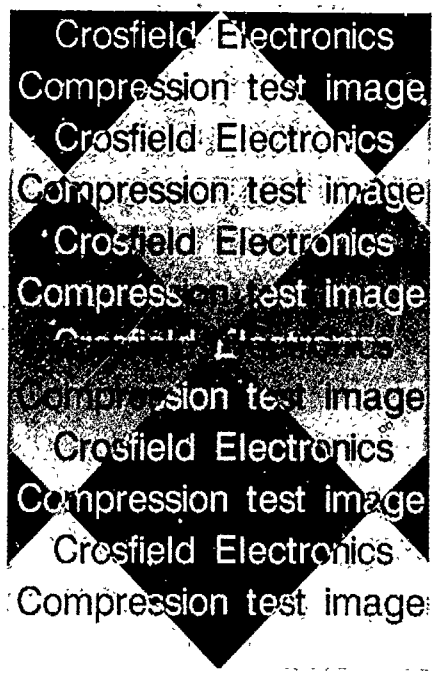


Fig. 2. Same image as fig. 1 after Histogram Segmentation Filtering



In a scan of a photograph such features are rarely a problem: a correctly sampled edge should not be sharp, and there is usually noise and texture present. We notice the worst effects in the artificial features of an assembled page such as text, borders, and cut-outs.

We can see these errors easily in the example in figure 1 whenever an 8*8 block overlaps a sharp edge. The original image had flat text, sharp edges, and smooth background. The compression was done using an experimental DCT algorithm that could compress scanned photographs with no visible degradation.

We could segment the blocks using the graph theory approach of Morris and Constantinides [5] to determine whether the block is likely to have been corrupted and to clean it up. Instead we have segmented the histogram. This gave much the same sort of result and was simple to implement.

2. SEGMENTING THE HISTOGRAM INTO PEAKS

We calculate the histogram of our 8*8 pixel block. We have a set of peak values V(n), n=1..N with frequency F(n) > 0. We may segment this histogram as follows...

(1). Find the two peaks with the closest values.
(2). If these peaks are separated by more than a threshold value, then stop.
(3). Merge the two peaks...
Frequency = sum of old frequencies
Value = weighted average of old values.
(4). Loop back to (1).

This should tidy the histogram up, sweeping loose clusters of values into sharp peaks.

3. THE SEGMENTATION THRESHOLD

When the threshold was kept constant the image lost its low contrast features, while high contrast noise from high contrast edges remained.

Making the threshold a fixed fraction of the total range in the block gave better results, but that was too easily influenced by extreme values.

The best results for this compression system were given by...

Threshold = Standard deviation * ß

This gave the same sort of filtering on a low and a high contrast block. This suited the compression system we were using. A value of ß = 0.6 gave a good balance between filtering noise and filtering out genuine image detail.

On low contrast blocks the calculation can be dominated by the rounding errors. To avoid this we turned off the filter if the standard deviation fell below 15.

4. DETECTING WHERE TO SEGMENT

We can determine what sort of image we have in our block from the number of peaks in our segmented histogram: Our test image (fig. 1) had only one peak for the flat blocks, and 2-4 peaks for the blocks with a feature. Typical scanned image blocks could have up to 9 peaks.

If the standard deviation was greater than the section 3 limit filter as follows:-

Peaks	Action
1	No filtering. Either flat or subtle texture. Compression should not have caused significant error.
2-4	Substitute corresponding segmented histogram values. Probably flat areas with sharp edges.
>5	No filtering. Probably texture.

5. RESULTS

The effects of the histogram segmentation filter are shown in Fig. 2. Most of the artefacts from compression have been removed.

The same filter was tried on filtered and unfiltered versions of real images containing text, colour charts, wood grain, face, and hair details. The results are not presented here. Only a few blocks were altered and no harmful effects were seen.

6. CONCLUSIONS

We have a filter that can remove fine detail from sharp edged features. This may be used to clean up the severe problems with text resulting from compression and decompression. As the filter acts on the image after decompression, it has no effect on the data compression ratio. It appears to have little effect on ordinary image data.

We can get a better value of threshold if we get more data from the compression system. We may estimate the quantization error to be zero if the quantized value is zero, and (quantization interval)/4 if not. All of the errors added in quadrature give an estimate of the error signal amplitude. We have used this approach with success [6], but it may not work with all compression systems.

[1] W.K.Pratt 'Digital Image Processing'
ISBN 0-471-01888-0 1978
[2] R.J.Clarke 'Transform Coding'
ISBN 0-12-175731-5 1985
[3] R.C.Gonzalez, P.Wintz
'Digital Image Processing'
ISBN 0-201-11026-1 1987
[4] JPEG Technical Specification
Source W.B.Pennebaker (IBM)
ISO/IEC JTC1/SC2/WG8 1990
[5] O.J.Morris, A.G.Constantinides
IPP Proc 133 F p146 1986
[6] R.A.Kirk British Patent application
8920905.0 1989

ANALYSIS IN PRINT PROCESS DEVELOPMENT: TWO CASE STUDIES

N. J. KERRY
Cambridge Consultants Ltd
Science Park
Milton Road
Cambridge CB4 4DW

Abstract - This paper describes two case-studies in which mathematical analysis and computer simulation have been used in the development of printing equipment and processes. The common theme of both examples is ensuring uniformly high image quality, although the printing technologies involved are very different.

The first example describes a design tool for offset litho print towers, and the second describes the analysis of acoustic effects in a novel multi-channel drop-on-demand ink jet printer array. Both case studies illustrate the way in which simple mathematics can be exploited to address complex and challenging design problems.

DESIGN TOOL FOR AN OFFSET LITHO PRESS

In the offset lithography process the image is etched onto the surface of metal plates. This etching changes the surface properties so that the plate becomes oleophilic. Ink is applied from one or more 'forme' rollers carrying a thin film of ink, which the image plate accepts only where is has been etched. The ink is then offset onto a blanket roller before final transfer to the paper or other medium.

The ink film on the forme roller is replenished from a supply via a series of intermediate rollers arranged in a tower. This tower must perform two functions. First, it must act as a buffer between the intermittent supply and the forme roller - it is not possible to provide adequate control of the replenishment with a continuous supply. The second function of the tower is to even out the variations in film thickness around the forme where ink has been transferred to the image. Poor performance by the tower will result in variations in overall image density, or in 'ghosting' within the image.

Generally speaking, the more rollers there are in the tower the better the performance - typically there might be 12 to 20 rollers in a small press. For a fixed-format machine in which the image size is standard, the tower configuration can be optimised by a process of trial and error starting from existing designs.

In order to design a variable format print tower, in which different image sizes are to be accommodated, a more sophisticated approach is needed. The requirement for a variable format makes is desirable that there is only a single forme roller, which makes is more difficult to achieve an even inking of the image. Furthermore, the tower configuration must be 'optimised' so that none of the image sizes would suffer from poor quality.

In order to enable designs to be evaluated on paper, a computer simulation was developed. This modelled the variation in ink film thickness around the circumference of each roller. A number of simplifying assumptions were made:

- the rollers are assumed to be in rolling contact without slippage;
- ink flow on the surface of the rollers is neglected, so that ink is assumed to be transported between the contact points around the perimeter of the rollers.
- the simulation was one-dimensional, so that longitudinal motion of the rollers was not modelled.

A simple ink transfer law was assumed at each point of contact - the total amount of ink prior to contact is divided between the rollers after contact, according to a fixed ratio.

The simulation is started from an initial fully inked state, and settles down into a steady state from which statistical measurements of the variation in image density can be obtained. It was validated qualitatively by comparing the predictions for existing towers known to give respectively good and poorer quality images, and then used to predict the performance of alternative designs and for fine tuning of the preferred candidate.

The use of this tool resulted in a roller configuration which required no further development before being put into production, and produces print of excellent quality, exceeding expectations.

CROSS-TALK EFFECTS IN HIGH DENSITY INK JET ARRAYS

Print quality in an ink jet printer is determined primarily be the accuracy of the landing positions of the drops on the paper. Errors in drop position can arise from a number of sources, and in this paper we are concerned with cross-talk - variations arising from interactions between one nozzle and its neighbours.

In designing an ink jet array for a given print quality requirement, the tolerable error in landing position must be shared carefully between these various sources. Typically, for a 300dpi printer, this overall tolerance is in the region of 20μm, and cross-talk effects must not exceed about ¼ of this.

The landing error depends on the relative error in drop velocity. It is highly desirable to operate at a drop velocity such that satellite drops do not form - about 3m/s for typical inks. For a high speed array printer, this brings the permissible variation of drop speed to around ¼m/sec (about an 8% variation) in order to keep the landing error to less than 5μm.

Analysis of the 'shared wall' actuator design

The figure below shows a novel design of printer array, fabricated from PZT, which can be made at densities up to 300dpi. In this design each wall is actuated in shear mode so that it deflects sideways and can operate the two channels either side of it.

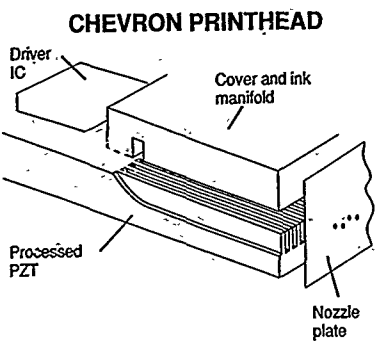


Fig. 1. Shared wall actuator, fabricated from a single block of PZT ceramic.

Although this design has many advantages, the walls are relatively compliant. This results in coupling of pressure between any channel and its near-neighbours, and is the primary source of cross-talk.

The volume of the delivered drop depends on the flow-rate through the nozzle and on the drop delivery time. The flowrate depends on the channel pressure generated, and the delivery time is just the time taken for an acoustic wave to propagate along the length of channel. The analysis described in this paper was initially undertaken to explain the apparent variation of acoustic velocity for different firing patterns in a prototype actuator. It has been developed specifically for this type of array, but is generally applicable.

The actuator is modelled as a number of identical two-dimensional channels containing ink. As described above, the walls separating the channels are compliant, and a pressure difference across the walls will cause a proportionate lateral deflection. We may neglect wall inertia as the resonant frequency of wall vibration is much higher than the frequencies associated with drop ejection.

The equations which describe flow in the channel are the momentum equation, the mass conservation equation, and the constitutive relationship between pressure and density for the ink. The flow is essentially one-dimensional and of small amplitude, and the linearised acoustic equations for each channel are:

$$\frac{\partial^2 p_i}{\partial t^2} + \kappa \frac{\partial^2}{\partial t^2}(p_{i+1} - 2p_i + p_{i-1}) = c_0^2 \frac{\partial^2 p_i}{\partial x^2} \quad (1)$$

Where κ is a constant, the ratio between the compliance of the wall and of the ink. The set of equations (1) can be written as a matrix wave equation:

$$(\underline{I} + \kappa \underline{A}) \frac{\partial^2 \underline{p}}{\partial t^2} = c_0^2 \frac{\partial^2 \underline{p}}{\partial x^2} \quad (2)$$

where $\underline{A}$ is the second-difference matrix. For an N-channel array there are N wavelike solutions $\underline{p}(x,t) = \underline{p}(x \pm ct)$ corresponding to the eigenvectors of the matrix $\underline{A}$. These the acoustic modes of the array, and for each eigenvalue λ the corresponding mode propagates non-dispersively with speed $c = c_0/\sqrt{(1+\lambda\kappa)}$.

The eigenvalues of $\underline{A}$ lie between 0 and 4, so that for a typical compliance ratio $\kappa = 0.3$ the theoretical variation in propagation speed is around 30% of c_0, consistent with the observed variability.

The theoretical analysis has been confirmed by comparison with experiment, and has shown excellent agreement. The existence of acoustic modes is confirmed by measurements of the resonant frequencies of an array with open ends. The expected linear relationship between eigenvalue and (period)^2 has been demonstrated, and agrees with independent measurements of the compliance ratio κ.

Cross-talk and cancellation

The analysis can be readily extended to give the channel pressure distribution $\underline{p}$ generated by a given excitation pattern $\underline{V}$:

$$(\underline{I} + \kappa \underline{A})\underline{p}_0 = \alpha \underline{V} \quad (3)$$

where α is the excitation parameter. This shows that when a firing voltage is applied to only a single channel, the effect of wall compliance is to reduce the pressure in the actuated channel and to induce cross-talk pressures throughout the array.

A video illustrating cross-talk in an actuator will be shown.

Equation (3) enables the pressure field generated by a given applied voltage pattern to be computed, but also offers a method of compensating for the effect of compliance. It is straightforward to determine the required initial pressure distribution to generate any given pattern of drops, and the above equation can then be solved, giving the voltage pattern needed to generate the required pressures.

In this way, with signal processing incorporated into the drive electronics, the array can be made to operate entirely free of cross-talk.

A GENERALISED MODEL OF LETTERPRESS INK TRANSFER

Roy Roach
Manchester Polytechnic
Faculty of Art and Design
Chatham Building
Manchester, M15 6BR
United Kingdom

ABSTRACT A general model is introduced for the transfer of ink to paper from solid letterpress plates. By isolating the non-linear from the linear region in the graph of ink transferred against ink available, attention is focussed on the need for a model giving the area of ink in contact with the paper at low ink film thicknesses. A possible model is proposed for this contact area.

INTRODUCTION

The experimental technique for investigating ink transfer requires the amount of ink (x), applied to the letterpress plate, to be varied and the resulting ink transferred (y) to be measured. This can be done by weighing the plate before and after the application of ink and following the transfer of the ink to the paper. From the resulting set of (x,y) co-ordinates, two graphs can be plotted. Figure 1 shows the first graph: ink transferred against ink available (y against x). Usually this has the typical "S" shape shown and the upper region which corresponds to high ink film thicknesses is normally linear. Figure 2 shows the second graph: proportion of ink transferred against ink available (y/x against x). This has a distinct maximum. It is this curve which has received most attention in modelling the phenomenon.

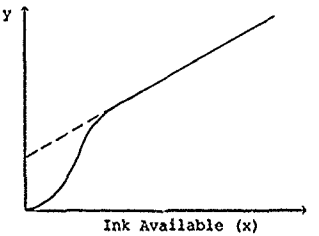


Figure 1

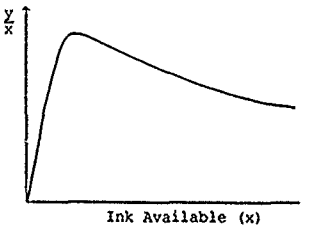


Figure 2

In their classic paper of 1955, Walker and Fetsko[3] laid down the foundations of the theory of ink transfer. Since then various criticisms have been made of the model they proposed but invariably other workers have established it as a basis and have attempted to improve upon it by minor modifications. Also it has been extended to the processes of gravure and lithography. For an excellent review and comparison of the various models proposed over the years for letterpress, the reader is referred to the 1982 paper by Mangin et al[2].

THE WALKER AND FETSKO MODEL

At very low ink film thicknesses, the contact between the paper and ink is incomplete due to the microscopically rough nature of paper. Walker and Fetsko suggested that the area in contact, A, be given by the exponential function

$$A = 1 - e^{-kx}$$

where k is a constant.

Focussing next on the paper in contact with the ink, Walker and Fetsko proposed that paper has a limiting capacity for "immobilising" or absorbing ink during the impression time. Again they suggested an exponential function to give the quantity of ink immobilised per unit area y_2

$$y_2 = (1 - e^{-x/b})b$$

where the constant b represents the maximum quantity of ink per unit area immobilised by the paper.

Consider next the ink which is not immobilised by the paper, i.e. the free ink film. Since the total amount of ink available for transfer, y_1, is given by $y_1 = x$, it follows that the quantity of free ink is given by $y_1 - y_2$. Walker and Fetsko proposed that the free ink film splits by a fraction f which is constant independent of x, so that the fraction of the free ink film transferred to the paper will be $f\cdot(y_1 - y_2)$.

Combining these relationships, it is seen that the quantity of ink transferred y is given by

$$y = A\left[y_2 + f(y_1 - y_2)\right]$$

$$y = (1 - e^{-kx})\left[(1 - e^{-x/b})\, b(1 - f) + fx\right]$$

At high ink film thicknesses, i.e. as $x \to \infty$,

$y = b(1 - f) + fx$.

a straight line with intercept $b(1 - f)$ and gradient f.

DEVELOPMENT OF GENERAL MODEL

The first assumption to be made here is that the printing plate is capable of immobilising a small quantity of the ink , say x_o. Define a new variable, z, the maximum amount of ink

available for transfer to the paper such that $z = x - x_o$ and $y_1 = z$.

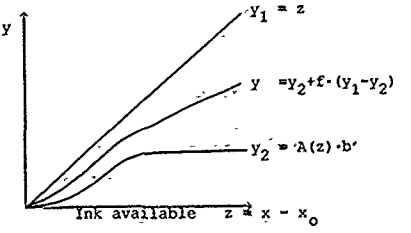


Figure 3

Suppose the maximum amount of ink immobilised by the paper is given by b', as for the Walker Fetsko model, but that the quantity of ink immobilised by the paper is given by

$$y_2 = A(x - x_o) \cdot b' = A(z) \cdot b'$$

where $A(x - x_o)$ is the fractional area of paper in contact with the ink and is some as yet unspecified function of $x - x_o$.

Now $A(z) = A(x - x_o)$ must have the following properties

As $x \to x_o$, $A(z) \to 0$

As $x \to \infty$, $A(z) \to 1$

The gradient of A(0) must not exceed unity

and A(z) has the familiar "S" shape.

Again adopting the Walker Fetsko technique, it is seen that the free ink film is given by $y_1 - y_2$ and assuming that a constant fraction "f" of this is transferred to the paper, independent of x it follows that the quantity of ink transferred

$$\begin{aligned} y &= y_2 + f \cdot (y_1 - y_2) \\ &= A(x - x_o) \cdot b' + f[(x - x_o) - A(x - x_o) \cdot b'] \\ &= A(x - x_o) \cdot b'(1 - f) + f(x - x_o) \end{aligned}$$

or $\underline{y = A(z)\, b'(1 - f) + fz}$

At higher ink film thicknesses $x \to \infty$ and A(z) is defined to approach 1 giving

$$\underline{y = b'(1 - f) + fz}$$

ISOLATION OF NON-LINEAR REGION

In order to focus attention on the type of function required for the contact area A(z), it is possible to extract the A(z) trend from practical data by isolating the non-linear region of the ink transferred against ink available graph (y against x). It is possible to determine for the linear region the gradient, f, and the intercept, b'(1-f), by a technique such as least squares. It is seen that

$$A(z) = \frac{y - fz}{b'(1 - f)}$$

If such calculations are made for all non-linear data points then a graph of A(z) against z can be constructed. This will give the S shape to be modelled by the function A(z). This curve lends itself to numerical techniques to determine the function.

CONTACT AREA MODEL

Several models were tried which have the necessary properties, defined for A(z). That which gave the best fit to practical data is based on the work of Bay-Sung Hsu[1], who investigated the variation of the area of contact of paper with the thickness of the ink film applied. Hsu proposed, that paper obeys a relationship of the form

$$\frac{\text{fractional area in contact}}{\text{fractional area untouched}} = \frac{A}{1-A} = Ct^n$$

A = fractional area of paper in contact with the ink

c and n are constants of paper

t = original thickness of the ink film measured from a plane through the tops of the maximum peaks

From this it follows that the fractional area of paper in contact with ink will be

$$A = \frac{Ct^n}{1 + Ct^n}$$

Hsu found this function to give a very good fit to experimental data measured by previous investigators, the value of n varying between 1.7 and 2.7 depending on the type of paper. The problem in this context is how to relate the ink film thickness below the maximum plane to the quantity of ink on the printing plate. It is assumed that the top peaks on the paper cannot penetrate the layer of ink of quantity x_o, immobilised by the plate. Accordingly the original ink film thickness below the maximum plane will be $z = x - x_o$ and the new model for the fractional area of contact becomes

$$A(z) = \frac{Cz^n}{1 + Cz^n}$$

Substituting into the general model gives

$$y = \frac{Cz^n}{1 + Cz^n} \cdot b'(1-f) + f \cdot z$$

and the proportion of ink transferred is given by

$$\frac{y}{z} = \frac{Cz^{n-1}}{1 + Cz^n} \cdot b'(1-f) + f$$

It follows that at the maximum, the z value and the area of contact are given by

$$z_m^n = \frac{n-1}{C} \quad \text{and} \quad A_m = \frac{n-1}{n}$$

BIBLIOGRAPHY

1. HSU, BAY-SUNG "Distribution of Depression in Paper Surface : A Method of Determination". Brit. J. Appl.Phys., Vol 13, 1962.

2. MANGIN, P J, LYNE, M B, PAGE, D H, DEGRÂCE, J H "Ink Transfer Equations - Parameter Estimation and Interpretation". APST, Vol 16, 1982.

3. WALKER, W C AND FETSKO, J M "A Concept of Ink Transfer in Printing". Am Ink Maker, 33, No 12, 38, 1955.

ELECTROTHERMAL RIBBON PRINTING PROCESS

RICHARD J.CAMERON AND DAVID N.M.IBRAHIM
Manchester Polytechnic
Chester St
Manchester M1 5GD UK

Abstract

A mathematical model of the electro-thermal ribbon printing process has been developed. The process and the model are described and some results of the heat-flow calculations are given.

Introduction

The electro-thermal ribbon (ETR) printing process was developed by IBM under the trademark QUIETWRITER. The process uses a matrix of dots and is capable of printing up to 200 characters per second with high quality. Ink on the ribbon is heated by heat generated electrically within the ribbon. The current enters the ribbon through a set of continuously moving electrodes. The paper and ribbon are in contact only for about 5 ms. The heat generated flows into the ink by conduction, the ink becomes sticky with the heat and adheres to the paper forming a black dot. A lot of early work is described in the IBM Journal of Research and Development, Volume 29 1985.

The aims of the modelling are to predict print quality from the material parameters and to improve control over the electric heating of the ribbon.

Structure of the Ribbon

The ribbon has three layers, as shown in Figure 1 and is traversed by an array of tungsten electrodes each of which is about 50 microns square with a spacing between electrodes of 50 microns. The polycarbonate gives the ribbon structural strength and is doped with carbon black to make it electrically conductive. The aluminium is sputtered on to the polycarbonate and acts as a return path for the electric current. The ink is applied either directly onto the to the aluminium or on to a very thin release layer.

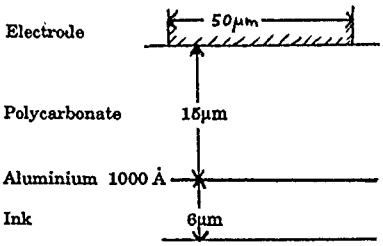


Figure 1. Structure of the ribbon

Heat Generation

Heat is generated in the ribbon by three distinct mechanisms :

i. in the body of the polycarbonate by Joule heating

ii. at the contact between the polycarbonate and the Tungsten contact

iii. at the boundary between the polycarbonate and the aluminium.

The last is due to water in the polycarbonate combining with the aluminium, during the sputtering, to produce a very thin, possibly monomolecular, layer of oxide. A significant proportion of total heat generated is generated in the oxide layer. This is beneficial to the process because heat is generated close to where it is needed, in the ink.

The contact and oxide resistances behave like Zenner diodes. Their break-down voltages introduce a non-linearity into the boundary conditions for current flow.

In order for the process to work, the ink must be subjected to very rapid heating which in turn requires that the energy densities in the heated region of the ribbon must be very high.

During normal operation the polycarbonate undergoes a glass transition at about 150C which requires the addition of energy. Below 150C it behaves like an ordinary resistance which of course varies with temperature. The current voltage characteristic of the ribbon is shown in Figure 2.

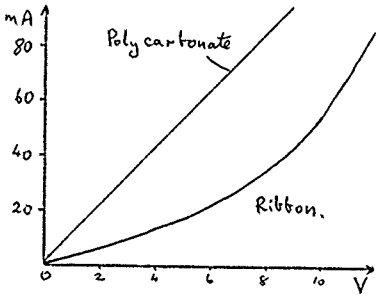


Figure 2 Current voltage characteristic of the ribbon.

The glass transition stresses the film with the result that the film buckles after printing and is usable only once. The changes in the state of the polycarbonate and of the ink make it necessary for the model to be able to take account of variation of material parameters with temperature.

Calculations

The current flow in the ribbon is calculated at the same time as the heat transport using finite differences with a variable grid. The current within the polycarbonate obeys the equation

$$\nabla.\sigma\nabla\varphi = 0$$

where φ is the electric potential and σ is the electrical conductivity.

The normal heat conduction equations are complicated by the glass transition in the polycarbonate and melting of the ink. Iterative techniques are used to establish i. the boundary conditions for current flow at the oxide layer and ii. the regions of the ribbon undergoing the glass transition.

Typically, currents of 25 mA are used in the printing at 40 characters per second, giving a potential difference across the ribbon of 7V. Figure 3 shows the pattern of the electric potential in a section across the ribbon below the electrode. In Figure 3 the electrode is at 7V and the aluminium is at ground. Figure 4 shows the temperature pattern after current has flowed in the ribbon for 0.5ms. The ambient temperature is 20C. It can be seen that the heat is produced largely under the 'footprint' of the electrode and that there is little interference betweeen adjacent electrodes. Figure 4 also shows clearly the importance of the aluminium oxide layer to the heating process.

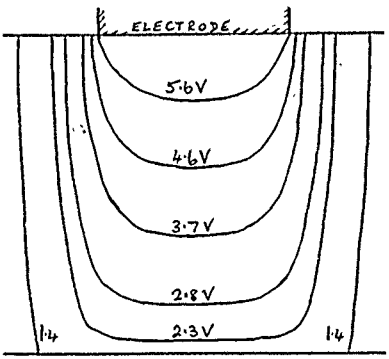


Figure 3. Electric potential in the ribbon

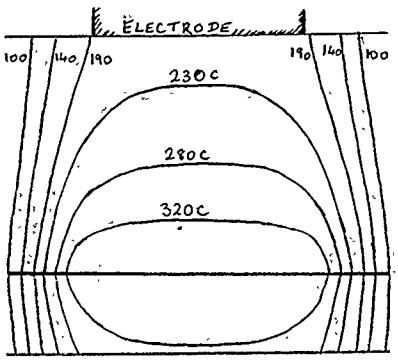


Figure 4. Heat pattern after 0.5ms

Validation

One of the major problems with modelling printing processes is calibrating and validating the model. Here, the small size of the components is compounded by the speed of the processes and the wide range of temperatures that are involved. Observation of temperature in the ribbon has only been possible by using a special printing rig with infra-red microscope observing the surface of the ink.

Acknowledgements

This work has been carried out as part of project 2125 of the ESPRIT initiative of the EC. We wish to thank our collaborators at Olivetti and AEG Olympia for their assistance.

A categorical approach to the semantics and implementation of discrete event simulation languages

M. Graña, M C. Hernández, A. d'Anjou,F.J. Torrealdea
Fac. Informática UPV/EHU, Aptdo 649, 20080 San Sebastián, Spain

Abstract: A semantic framework for discrete event languages has been defined and built up. Its use as a tool for abstract description and implementation of such languages is sketched.

Semantic description of discrete event languages

Discrete event languages have been traditionally presented by means of a series of examples. When needed, a summary account of the inner operation of the simulation mechanism is provided, usually to prevent model misbehavior due to this very mechanism. On the other hand, people planning to build some specific purpose (restricted domain) simulator usually must reason in very primitive terms, for lack of a proper level of abstraction. In both cases what is missed is a semantic context to describe precisely the meaning of the langage constructs. Given a proper semantic framework, the rigorous and complete description of the behavior of each language construct could be provided or designed in a very compact form.

The category of system specifications

The formal setting we are thinking of is based upon the formal developments of Zeigler [4]. The hierarchy of system specifications originally proposed in [4] can be assumed as the domains of a semantic category which provides the meaning of discrete event languages The structure of such a category is given in fig. 1. The arrows in this figure represent two kinds of functions: semantic (sf_i) and behavior (b_i) funtions. Semantic functions provide the interpretations of objects in a domain as objects in a lower domain. Behavior functions map any other domain into the distinguished domain DS, which specifies the collection of traces and statistics that constitute the behavior of systems. So, behavior functions provide the behavior associated with systems specifications. The hierarchy of system specifications is a hierarchy of operational abstractions and the category of system specifications is the formal interpretation of those abstractions.

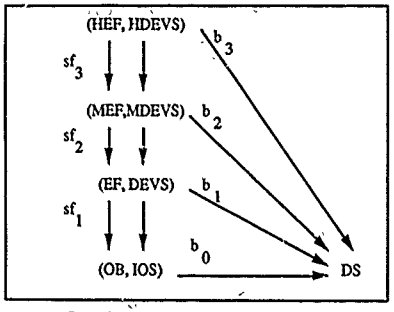


figure 1 The system specification category

Objects in the IOS domain correspond to Input/Output systems whose specification includes their input, state and output spaces, along with the transition and output functions. The behavior function b_0 of this domain involves the iterative application of those functions.

Objects in the DEVS domain correspond to the discrete event specifications originally proposedby Zeigler. The semantic function sf_1 involves the composition of the internal and external transition functions with the time advance function to construct the transition function of the equivalent IOS object. The behavior function $b_1$can be computed as $b_0 \circ sf_1$.

Objects in MDEVS and HDEVS domains correspond, respecuvely, to the modular and hierchical specifications. Their respective semantic functions sf_2 and sf_3 provide respectively the mechanisms to couple components and submodels. Again, the respective behavior functions can be computed as compositions of semantic functions, that is.

$$b_2 = b_0 \circ sf_1 \circ sf_2$$
$$b_3 = b_0 \circ sf_1 \circ sf_2 \circ sf_3$$

Objects in the OB, EF, MEF and HEF domains correspond to the specification of experimental conditions at the various levels of the hierarchy, that allow the selection of subsets of the complete behavior of the system (empty experimental frame). The semantic functions apply also to them.

Implementation of the semantic framework

A concrete specification of the domains and functions involved int this category of system specifications was given in [1]. This specification can be straightforwardly implemented using any object oriented programming language. In fact, it has been implemented using SIMULA [2]. In brief, domains are implemented as classes whose bodies realize the semantic functions over their respective domains. The structure of arrows in fig.1 is implemented, then, through the class inheritance structure. Finally, the mechanism of virtual function definitions are used to allow any final system transition functions to be implemented. The user of this implementation of the semantic framework can abstract from any operational consideration and concentrate on the specification of the particular mechanics of his language.He does so specifying his language constructs as MDEVS or HDEVS objects.

Our approach diverges from the implementation of the DEVS formalism reported in [5] in the sense that the semantic framework is not intended to be a language by itself, but as a mean to conceptualize and realize more specific languages.

Applications of the semantic framework

The framework has already been used to implement a subset of RESQ [3], including its hierarchical features. It is also being used as teaching support for a simulation subject. The students use it to implement a simulator for generalized stochatic Petri Nets. Other languages, dialects of GPSS and SLAM (including MHEX) are under analysis and semantic specification headed towards their implementation upon the semantic framework. The final goal is to get a collection of languages such that for each of them a rigorous semantic description has been worked out, and an implemetation following it has been realized.

Extensions to the semantic framework

The inclussion whithin the formal framework of concepts relative to the graphical presentation of the model's behavior and state could be of great interest in order to provide the semantics for languages that include animation capabilities.

Semantic specifications of the languages must be independent of the concrete computational device (sequential or distributed), they are operationally independent. Distributed realization of the semantics of the upper semantic domains means the distributed realization of the languages mapped into them. This is obviously independent of the precise distributed strategy (whether conservative or optimistic). We are also searching in this direction.

References

[1] Graña M. 1989 Una contribución a la especificación formal de los lenguajes de simulación discretos Tesis Doct. Fac. Informática UPV/EHU
[2] Hernandez M.C., M. Graña 1991 Implementación de una base semántica para lenguajes de simulación de sucesos discretos Int Rep. Fac. Informática UPV/EHU
[3] McNair E.A., Sauer C.H. 1985 Elements of practical performance modelling Prentice Hall
[4] B.P. Zeigler 1976 Theory of Modelling and Simulation Wiley
[5] B.P. Zeigler 1987 Hierarchical, modular discrete event modelling in an object oriented environment Simulation 49(5) pp 219-230

Aproximate linear complexity resolution of the Satisfiability problem via Boltzman Machines

A. d'Anjou, M. Graña, F.J. Torrealdea, M.C. Hernandez
Fac. Informática UPV/EHU, Apdo 649, 20080 San Sebastián, Spain

Abstract: A series of experiments have been performed applying Boltzmann Machines (BM) to the resolution of the Satisfiability (SAT) problem in the propositional calculus setting, which shows some interesting features: resolution time insensitivity to clause type and size, and linearity against the number of propositions involved.

BM and the SAT problem

BM's are a class of recurrent neural computing mechanisms that have been proposed originally in [3]. BM's are defined by giving a set of logical units, connections between them and the strength associated with each connection. Computing in the BM is performed through search of the max/min consensus configuration by the simulated annealing algorithm. A configuration of the BM is a map from units to their local states. The consensus function gives a consensus value for each configuration which is computed as the summation of the strengths associated with the set of active connections in the configuration. A connection is active when the states of all the units connected by it are "ON". A special kind of unit, the sigma-pi unit (introduced in [5]) serves to model higher order connections between more than two units, allowing higher order expressions for the consensus function. A broad discussion of the application of BM's to optimization, classification and learning can be found in [1].

On the other hand, the SAT problem is a well known problem in computer science [2]. It can be stated as a particularization of the more general MAX-SAT problem (finding the maximum subset of clauses satisfiable simultaneously), where the maximum searched is the whole set of clauses. Under this interpretation of the SAT problem we dared to apply BM to its resolution. As already proved by Pinkas [4], any propositional SAT problem can be mapped into a recurrent neural network. He also showed how sigma-pi units can be introduced, reducing the number of hidden units involved, and increasing the order of the energy function (the consensus function in BM).

Our approach has been a straighforward one. The clauses have been modelled by variable order (the size of the clause) sigma-pi units, and the propositions by common (order two) units. We built up a BM to model a given SAT instance, and a maximum consensus configuration is searched. The SAT instance is positively answered when a configuration in which all the clauses are satisfied is reached. As a side product, a random truth assignement that satisfies them is obtained. The negative answer is produced when a number of trials (annealings) have been performed without success.

Experimental results

Through the experiments reported here, a quite conservative annealing schedule has been used, because we were more interested in the cualitative features of the approach, and it was our intention to factor out the effects of poor annealing strategies. This accounts for the high values of the performance metric shown in the figures. The basic performance metric (time in the figures) is the number of trials in the annealing that reaches a satisfaction configuration for a given instance. We realized the experiments over sets of 30 instances generated randomly along some parameters: size of the clause (Uniformly distributed,2,5,6,7), number of clauses (NC) and number of propositions (NP).

The figures include 95% confidence intervals and interpolation lines. Figure 1 shows that for clause sizes distributed uniformly up to 7 propositions per clause, the time needed to reach the satisfaction configuration grows linearly with the number of propositions involved. figure 2 and figure 3 show, respectively, that the time of resolution is insentitive to the distribution of clause sizes and the number of clauses.

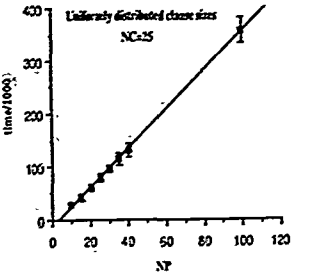


Figure 1 Linear behavior of time versus NP

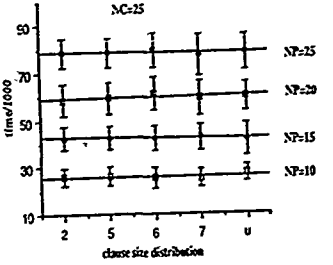


Figure 2 Independence of time from the distribution of clause sizes

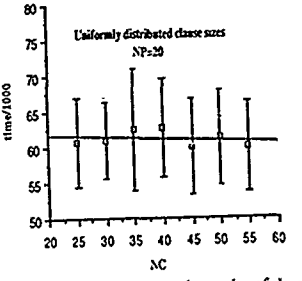


Figure 3 Insensitivity to the number of clauses

References

[1] Aarts E.H.L., J H M. Korst Simulated annealing and Boltzmann Machines Wiley Chichester 1988
[2] Garay M R., D.S. Johnson Computers and intractability W.H Freeman San Francisco 1979
[3] Hinton G E, T J Sejnowski, D.H. Ackley Boltzmann Machines: constraint satisfaction machines that learn Tech Rep CMU-CS-84-119 Carnegie Mellon Univ.
[4] Pinkas G Energy minimization and the satisfiability of propositional logic Proc. 1990 Connectionist Models Summer School. Morgan Kauffman San Mateo CA
[5] Sejnowski T.J. Higher-order Boltzmann Machines in J.S. Denker (ed) Neural Networks for Computing AIP 1986

FUNCTION APPROXIMATION ALGORITHMS FOR SYSTEM SIMULATION AND MODELLING

Z. JACYNO
Department of Physics
University of Quebec at Montreal
P. O. Box 8888, Station A, Montreal, P. Q.
Canada H3C 3P8

Abstract. A method of algorithm development for dynamical systems based on input and output signals decomposition with respect to basis in the functional Hilbert space is proposed. Application of the projection theorem permits the advance evaluation of the precision error and computing efficiency of the algorithms. A library of basis allows for their most suitable choice for a particular system, further enhancing the algorithms precision and efficiency.

I. INTRODUCTION

In the modelling and simulation of systems, the first arising problem is a good comprehension of their qualitative functionning followed by the quantitative description, both validated with respect to some criteria in the reference to a real system. As a result, a mathematical model is obtained which is then implemented on a computer, most often nowadays, a microcomputer. The mathematical model must undergo transformations into a "best" algorithm, meeting the needs and expectations of a design or field engineer.

The development of mathematical models still remains a scientific art, relying on accumulated knowledge in a particular domain. The development of the algorithms, though, may be put into a more defined and generalized framework. The purpose of this paper is to propose an approach which appears promissing in both the efficiency and precision of the resulting algorithms.

The mathematical model, in order to yield a computer algorithm, must undergo some mathematical processing such as time discretization, approximation of nonlinearities, etc. Currently, such processing relies mostly on the use of polynomials and is based on the Weierstrass approximation theorem. It hardly allows for an advance precision evaluation, neither does it assure the algorithms' convergence. This paper introduces higher order approximations by functions in the Hilbert space, in conjunction with the projection theorem for evaluation of the approximation precision.

II. OPTIMAL MEAN SQUARE APPROXIMATION USING FUNCTION BASIS

The projection theorem, [1, 2], needs reformulation for our purposes into a more compact matrix form. Real systems and their models when used in design or simulation are subjected to different kind of input signals. They are deterministic for testing and stochastic in control applications. The implementation of models on computer requires a representation of these signals in numerical form, meaning their approximation. We shall consider approximations by a set of linearly independent functions $\{\varphi_i(t)\}$ in the n-dimensional Hilbert space $L^2(t_1,t_2)$ with the inner product defined over it by

$$(\varphi_i,\varphi_j) = \int_{t_1}^{t_2} \varphi_i(t)\,\varphi_j(t)\,dt. \tag{1}$$

The L^2 space has the induced norm

$$\|f(t)\| = \left(\int_{t_1}^{t_2} |f(t)|^2 dt\right)^{\frac{1}{2}}. \tag{2}$$

Any signal x(t) may then be approximated by a linear combination of basis $\{\varphi_i\}$

$$\tilde{x}(t) = \sum_{i=1}^{n} a_i\varphi_i(t) = a^T\varphi(t), \tag{3}$$

where

$$\varphi^T(t) = \varphi^T = [\varphi_1\ \varphi_2\ \dots\ \varphi_n] \tag{4}$$

is the basis vector, and

$$a^T = [a_1\ \ a_2\ \ \dots\ a_n] \tag{5}$$

is the vector of approximation coefficients, suitably chosen.

The approximation (3) describes the initial signal x(t) with some error

$$e(t) = x(t) - \tilde{x}(t), \tag{6}$$

the value of which may be estimated by the norm (2) applied to (6) The value of the mean square approximation error is

$$\|e(t)\|^2 = \|x(t) - a^T\varphi(t)\|^2 \tag{7}$$

or

$$\|e\|^2 = \|x\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j(\varphi_i,\varphi_j) - \sum_{i=1}^{n} a_i(x,\varphi_i), \tag{8}$$

when properties of the inner product are taken into account. In short hand matrix notation, (8) is given by

$$\|e\|^2 = \|x\|^2 + (a^T\varphi, a^T\varphi) - (x, a^T\varphi), \tag{9}$$

or by

$$\|e\|^2 = \|x\|^2 + \varphi^T a a^T\varphi - x\varphi^T a. \tag{10}$$

In the approximation formula (3), the coefficients a

may be chosen to provide the best approximation possible in the basis φ. They must minimize the relation (9), which requires that the first variation

$$\delta(\|e\|^2) = 2(\varphi\varphi^T)a - 2(x,\varphi) \quad (11)$$

vanishes. It occurs when

$$\Phi a = B, \quad (12)$$

where the square nxn matrix Φ has the entries given by

$$\varphi_{ij} = (\varphi_i,\varphi_j) \quad (13)$$

and the column nx1 vector B, by

$$b_i = (f,\varphi_i). \quad (14)$$

The second variation of (9) indicates that the solution (12) indeed minimizes the approximation error (6) with respect to its mean square value.

Under condition (12) the minimal mean square error is then found to be

$$\|e\|^2_{min} = \|x\|^2 - a^T\Phi a. \quad (15)$$

It steadily decreases with the increase of the approximation order, i. e. with the increase of the dimension of the coefficient vector.

Equation (15) may also be developed from geometric relations between the approximation error and basis in the Hilbert space. In order to assure the minimization of the error, these vectors should be orthogonal

$$(e,\varphi) = 0. \quad (16)$$

This is called the projection theorem, [1], and it also yields the results given by (12) and (15).

The effects of the projection theorem are shown in Fig. below, where the approximation of the ramp function in term of exponential basis is illustrated

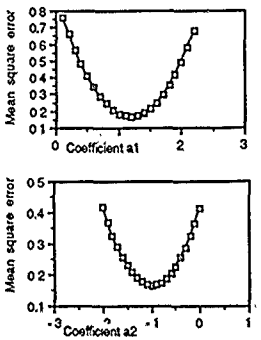


Fig. Mean square error for a unitary ramp signal approximation by the exponential basis $\varphi_i = e^{-it}$ and two-dimensional coefficient vector with the optimal value, (12), a^T=[1.19232, -0.99576].

III. MODELLING AND SIMULATION OF DYNAMIC SYSTEMS VIA FUNCTION APPROXIMATION

For dynamic systems, input and output signals are decomposed with respect to the same set of basis. Both signals and their approximations are related one to another through the mathematical model of a given system. Consequently, the chosen basis should represent a complete and closed set with respect to mathematical operators of the model. For linear systems, these operators are integrators and multipliers by a constant. Mathematical models of systems will be reflected by relationships between the input and output approximation coefficient vectors and given by integration matrices and the systems' coefficients only. Such important features as precision, convergence and efficiency of the algorithms could then be evaluated in advance from an imposed desired value of the approximation errors.

A library of different types of basis may be constructed allowing for the most suitable choice for a particular problem. This further and significantly enhances the efficiency of the algorithms. Indeed, it has been noted in litterature, [3], the many fold reduction of computing time in certain situations. Simultaneously, the algorithms' precision may also be increased since the approximation error (7) depends upon the choice of basis.

IV.CONCLUSIONS

The proposed approach of the algorithms' construction using function approximations offers good potentials for the increase of their precision and efficiency. Function basis, especially those with orthogonal or, better still, orthonormal properties, provide promissing new mathematical tools for the computer modelling and simulation of systems.

REFERENCES

[1] De Coulon, F., Signal Theory and Processing, Artech House, Inc., Dedham (Mass.), 1986.

[2] Keener, J.P., Principles of Applied Mathematics-Transformation and Approximation, Addison-Wesley Publishing Co., Redwood City (Ca.), Menlo Park (Ca.), Reading (Mass.), New York, Amsterdam, DonMills (Ont.), Sydney, Bonn, Madrid, Singapore, Tokyo, San Juan, Wokingham (U.K.), 1988.

[3] Cavin R. K., Wolff P. K., Palusinski O. A., Guarini M. W., Lee A. and Su Y., Efficient Method for Simulating MOS Integrated Circuits, Final Report, SRC Contract No. 83-01-038, Department of Electrical and Computer Engineering, The University of Arizona, Tucson, Arizona, 1987.

ON THE OPTIMAL ALGORITHM OF SIMULATION MODEL OF DEDS

Dai,Qinglin And Wang,Zhengzhong
Beijing Institute of Information and Control
P.O.Box 842, Beijing, PRC 100037

Abstract—This paper tries to explore the approach to solve optimization of DEDS simulation model and proposes a heuristic algorithm that is specifically designed to the DEDS simulation. Its basic idea is to approximate the optimal point gradually in the processes of iteration using a set consisted of several points instead of using one point as in conventional optimal techniques. Obviously such set can filter more stochastic errors caused by the simulation itself than one point can. Finally the paper testifies the algorithm on a typical DEDS example and the results show very satisfactory.

Keywords: Discrete Event,System Simulation, Optimization of System,Set Approximatting Algorithm(SAA)

1.Introduction

Discrete Event Dynamic Systems (DEDS) is a kind of complicated systems. Such systems as computer networks, transport and manufactory systems are vivid DEDS examples In general Analytical method and Simulation method are two main methods to study DEDS(Reference 1) Here only Simulation method is discussed.

At first the optimal problem of DEDS simulation discussed in this paper is described. For simplicity, the optimal problem is defined as:

$$\min \quad f(\theta) \qquad (*)$$

Here $f(\theta)$ is the object function of the simulation model of m-dimension decision variable $\theta=(\theta_1,\cdots,\theta_m)$ It is only after running the simulation model in computer that the value of $f(\theta)$ can be obtained. The object of problem (*) is to select suitable parameters θ^* and reach the minimum of $f(\theta^*)$ Problem (*) is the simplest form of all optimal problems. As for other forms of optimal problems of DEDS such as the problems with constrained conditions, further research is needed although the principle of the algorithm discussed here will be useful.

Up to now, there are no any mature methods which can manipulate optimal problem (*) easily. In general,several reasons listed below hinder the development of optimal techniques about DEDS simulation.

(1). The object $f(\theta)$ usually cannot be manifested analytically and even no consice and beautiful models to describe DEDS now. So it is difficult to study DEDS theoretically and there are almost no any general conclusions on DEDS simulation.

(2). Because the values of $f(\theta)$ obtained are simulation results of some random variable functions and which errors cannot be neglected, it is not probable to use conventional non-linear programming techniques on DEDS directly(Reference 4,5).

(3). It is rather difficult to compute the gradient values of $f(\theta)$ in DEDS simulation. In one hand a large number of simulation time is needed(especially for complicated multi-variable systems) and on the other hand the values of gradient got in simulation usually have big errors. So the information about gradients which plays an important role in conventional optimal process cannot be simply used in DEDS simulation areas.

In spite of the obstacles mentioned above, because of the importance of the optimal techniques of DEDS simulation, many researchers have already done a lot of work and made much progress although it still has very long way to go to be able to call them practical(Reference 2,3,6,7)

2. Set Approximatting Algorithm(SAA)

This section describe a new heuristic optimal algorithm about DEDS simulation.

At first a new concept is put forward which is in fact a set denoted by $\Phi(f)$. Here f stands for object function $f(\theta)$ and l for the number of elements in the set $\Phi(f)$. The set is consisted of l values $f(\theta_1),\cdots,f(\theta_l)$ of object function on l points $\theta_1,\cdots,\theta_l$.

Intuitively a big and heavy moving object cannot be influenced easily by random factors in environment and in the same reason that a set consisted of several values can filter more stochastic errors caused by simulation itself than single value can and thus corresponding algorithm will to the optimal point more accurately and smoothly than others do. This is just the heuristic principle of the algorithm proposed below and so this algorithm is called Set Approximatting(SA) algorithm.

Until now there is no any general conclusion about how to select l. In general, the bigger l is, the more accurate the optimal solution is and correspondingly the more simulation time will cost Here exists a trade-off between the accurate and the time. This problem is certainly important to the optimal theory of DEDS simulation and worthy careful exploring. Here temporiary let l=m+1, and m is the dimension of decision variable.

Below SA algorithm is described in detail and at first some denotations must be defined:

$$\min\Phi(f)=\min\{f(\theta_1),\cdots,f(\theta_l)\} \qquad (2.1)$$

$$\max\Phi(f)=\max\{f(\theta_1),\cdots,f(\theta_l)\} \qquad (2.2)$$

$$d\Phi(f)=\max\Phi(f)-\min\Phi(f) \qquad (2.3)$$

And then defining θ relatived to $\min\Phi(f)$ as θ_{min} and $\max\Phi(f)$ as θ_{max} here is also called the "worst" point in $\Phi(f)$. It will be replaced by better one according to the algorithm The improvement of the relation between the set and the optimal point will guarantee the algorithm to reach the optimal point gradually. The value of $d\Phi(f)$ is as the stopping criterion of the algorithm. When its value is very small that also means that the value of $\max\Phi(f)$ nears $\min\Phi(f)$ closely, at this time it can be said that the algorithm has already gotten satisfactory results.

SA algorithm can be described as below:

1). Initialize the set $\Phi^0(f)$ and let k=1;

2). Let $\Phi^{k+1}(f)=\vartheta[\Phi^k(f)]$;

3). If $d\Phi^{k+1}(f)<\epsilon$, then stop else goto 2).

(end of algorithm)

In the algorithm, ϵ is a positive constant, ϑ is the map from R^l to R^l. This map should guarantee the behavior of $\Phi^{k+1}(f)$ is better than that of $\Phi^k(f)$. That is just the condition below:

$$\min\Phi^{k+1}(f)<\min\Phi^k(f) \qquad (2.4)$$

$$\text{or } d\Phi^{k+1}(f)<d\Phi^k(f) \qquad (2.5)$$

In general the key to construct one succesful algorithm is to construct suitable map that will be discussed below.

The most important element in map of SA algorithm is how to get a new point from previous sets. Recently a new approach proposed by Y.C.Ho et al. put forward a convenient and face-to-simulation way that is called Perturbation Analysis (PA)method. Its main advantage is to be able to get the gradient of Performance Measure of the system with respect to parameter using only one Monte-Carlo experiment. Although PA method is not well-developed now and still needs a lot of work to do, the SA algorithm proposed here will use PA method to look for the search direction to the new point (Reference 8,9,10)

Mapσ is constructed as below:

At first obtaining a new point θ' using PA method, then getting $f(\theta')$ by simulation. This time there will occur one of three cases listed below:

1) $f(\theta') < \min_{i} f(\theta_i)$ (2.6)

2) $\min_{i} f(\theta_i) < f(\theta') < \max_{i} f(\theta_i)$ (2.7)

3) $f(\theta') > \max_{i} f(\theta_i)$ (2.8)

SA algorithm will replace θ_{worst} (worst point) with θ' (better one) if case 1) or 2) occurs and other components remain constant, $\{\theta_i\}$ is constructed and then obviously it will satisfy condition (2.4) or (2.5).

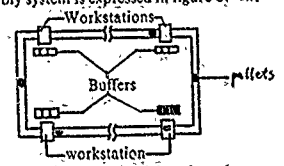
But as the value of $f(\theta')$ got in simulation unavoidable has error, it maybe cause the failure of the algorithm. For example, the simulation value of $f(\theta')$ is probably much less than real counterpart and maybe even less than accurate optimal solution of $f(\theta^*)$. If the occurence of such cases in simulation is more than 1, it is probably that all elements in set $\{\theta_i\}$ are less than $f(\theta^*)$ after Nth iteration. At this time only case(2.8) is occured and the algorithm cannot proceed in fact to the optimal point. So the real optimal solution cannot be obtained. In the operation of SA algorithm, more simulation length and time are used to decrease the occurence of such poor cases and thus the probablity of the failure of the algorithm is much less.

The stopping criterion chosed in SA algorithm is also suitable to the simulation, which can decrease the effect of random errors inherently with the simulation.

3.Case study

The effect of SA algorithm is testified on a typical DEDS(i.e an assembly system)

The assembly system is expressed in figure below:



Assumming the number of workstations in system is two, the service times of workstations are subject to exponential distribution and the mean service time are respevtively θ_1 and θ_2, the number of pallets in the system is eight. And there is a buffer behind the workstation 1 which item is two. In any time, the number of parts processed in system is constant(=8)

The object function of the system is defined as:

$$f = c_0 (T/N) + \sum_{i=1}^{2} (c_i/\theta_i)$$

Here $T = T(\theta_1, \theta_2)$ is the lenghth of the simulation time at each iteration, N is the number of parts processed in the period of time T, c_0, c_1, c_2 are constant numbers relative to the cost of the system, here let $c_0 = 1$ and $c_1 = c_2 = c$.

T/N in the object function is the average processing time of each part through the assembly system. It is required that the less the better; And $\sum_{i=1}^{2} (c_i/\theta_i)$ is the costs of the system itself(for simplify only the costs of workstations are considered) also the less the better obviously. So there exists a trade-off between the effiency and cost. For example, when θ_i ($i = 1,2$) is smaller then T/N is also smaller but this time more cost is required and vice-versa. The aim of system optimization is then to chose suitable parameter θ_1^* and θ_2^* minimize the object function f.

Because the system above has no analytical optimal solution, here one method which has alredy experimented successfully by Professor R.Sury et al (See Reference 13,14) is applied to compare with the results obtained by SA algorithm, this method is called method 3 here.

By using different sample of random series the SA algorithm can be divided into two suber-methods. Method 1 uses a variance reduction technique(CRN) So the same pseudo-random number generators are called and the same pseudo-random number series are used. So there are enough reasons to believe that the difference of any obsevable data is caused by the difference of system design instead of experimental condition. Method 2 try to dynamically update the system parameters in the process of simulation. So the random series used in each iteration is one section of a long stochastic series. This method is especially suitable for on-line simulation.

In experiment, the simulation model is written in GPSS-F simulation language and run in IBM-PC/AT(Reference 11,12)

After analyzing the results of the experiments, it is showed that the results obtained by SA algorithm are nearer the optimal point than those by method 3. Moreover from the experimentes, it is showed that in the neighbor areas of the optimal points, the optimal processes using method 3 are badly vibrate but those using SAA can approximate optiaml points smoothly.

4.Conclusion and Future Developments

This paper proposed a new heuristic algorithm which is specifically designed to DEDS simulation. The algorithm is called Set Approximatting (SA) algorithm. It is revealed that this method can get better optimal results than some other successful algorithms.

In the future, some further research work is needed such as to theorize the SA algorithm and improve the algorithm itself and apply the SA algorithm to more real complicated DEDS and so on.

Reference

1. Y.C.Ho(Editor) SPEEDS——A New Techniques of the Analysis and Optimization of Queueing Networks, Harvard University,Division of Applied Sciences, Technical Report No.675,feb. 1983.

2. F.Azadivar and Y.H.Lee (1988) Optimization of Discrete Variable Stochastic Systems by Computer Simulation, Mathematics and Computers in Simulation 30, 331-345.

3. Azadivar,F. and Talavage,J. (1980) Optimization of Stochastic Simulation Models, Mathematics and Computation in Simulation XXII, 231-241.

4. Avriel,M.(1975) Non-linear Programmming: Analysis and Methods, Prentice-Hall, Englewood Cliffs, New Jersey.

5. Deng,Naiyang,Computation methods of non-constrained optimization problem, Science press,P.R.China.

6. M.X.Meketon (1987) Optimization in Simulation: A Survey of Recent Results, Proceedings of the 1987 Winter Simulation Conference.

7. Glynn,P. E. (1986) Stochastic Approximation for Monte-Carlo Optimization, Proceedings of the 1986 Winter Simulation Conference, 356-365.

8. R.Sury (1989) Perturnation Analysis. The State of the Art and Research Issues Explained via the GI/G/1 Queue. Proceedings of the IEEE, Vol.77,No.1.

9. R.Sury (1987) Infinitesimal Perturbation Analysis for DEDSs, Journal of the ACM, Vol.14, No.3, 686-717.

10. Y.C.Ho (1987) Performance Evaluation and Perturbation Analysis of DEDSs, IEEE Transactions on Automatic Control, Vol. AC-32, No.7.

11. B.Schmidt. GPSS-Fortran, John Wileysons.

12. J.O.Henrikson State-of-the-art GPSS.

13. R.Sury and Y.T.Leung (1989) Single Run Optimization of Discrete Event Simulations--An Empirical Study Using the M/M/1 Queue. IIE Transactions, Vol 21, No.1, 35-49.

14. R.Sury and Y.T.Leung (1987) Single Run Optimization of a SIMAN Model for Closed Loop Flexible Assembly Systems. Proceedings of the 1987 Winter Simulation Conference, 738-748.

An Intelligent Modelling Interface to Numerical Simulation

N.J. Hurley, D.P. Finn and N. Sagawa
Hitachi Dublin Laboratory
O'Reilly Institute,
Trinity College,
Dublin 2,
Ireland.

Tel: +353-1-6798911
Fax: +353-1-6798926
Email: NHURLEY@VAX1.TCD.IE

Abstract
This paper presents the conceptual design of an intelligent interface to numerical simulation, which deals with the modelling of problems described by partial differential equations (PDEs). The complete modelling process from a given real world problem, to a simulation code is analysed and three intermediate models are extracted. These models correspond to engineering, mathematical and numerical representations of the problem. The system structure incorporates an entry level for each model and provides transformations between them. This allows users to specify the problem in the terminology of the model most suitable to their own expertise. A frame based approach to knowledge representation reflects the hierarchical nature of domain knowledge. Concepts which form the building blocks of each model are stored as frames in three knowledge bases. Model transformations are achieved by accessing localised rule bases bound to each frame.

1. Introduction

Numerical simulation has become a very important and powerful technique for scientific experimentation and design. As well as numerical analysts, many different kinds of people, including design engineers, experimental physicists and applied mathematicians can find use for it in their everyday work. With the increasing availability of powerful computer resources, there is a genuine need to provide numerical simulation software tools which can facilitate these potential users.

To design such a tool, the issue of model comprehension must be addressed. It is essential to cater for how users think about and express their problems. Existing numerical simulation packages such as DEQSOL [Umetani *et al.* 1985, Kon'no *et al.* 1986], ELLPACK [Rice 1985] and FIDISOL [Schonauer and Schnepf 1987], have greatly reduced the effort demanded by numerical simulation, compared with direct coding in FORTRAN or C, by providing high-level programming languages or drivers of subroutine libraries. However, without the prerequisite knowledge from the mathematical and numerical analysis domains, the use of these software tools can be formidably difficult. In other words, there is still a considerable conceptual gap between the input level of these packages and the understanding of a large proportion of potential users. Also, while offering quite a lot of flexibility to the knowledgeable user, they fail to provide sufficient guidance to the non-expert towards finding the best model, from a potentially large space of possibilities.

Several research projects are investigating the application of AI techniques to the area of numerical simulation. The Numerical Algorithms Group (NAG) [Chelsom *et al.* 1990] is developing knowledge-assisted numerical routine selection tools for the diverse NAG FORTRAN library. The EVE system is aimed at mathematicians [Barras *et al.* 1990] and enables users to create PDEs from pre-defined primitive mathematical components. Another system which makes use of PDEs as the interface level was reported by [Russo *et al* 1987], in which some of the numerical stability and efficiency constraints are taken into account. These projects mainly aim at users who are knowledgeable enough to make the right choices during the decision process or who are to a certain extent familiar with mathematical expression. However, little work has been done to establish a complete modelling framework, which incorporates the comprehension levels of all potential users, from engineers with no numerical or mathematical knowledge, through to numerical analysis specialists.

This paper addresses the conceptual design of a sophisticated tool for numerical modelling. Specifically, it is concerned with the modelling of problems which can be described mathematically by a set of PDEs. These include heat transfer, fluid dynamics, structural analysis and electric field analysis problems. The complete modelling process, from a given real world problem to the final simulation code, is examined, and three intermediate conceptual models are isolated. Each model can be considered as a valid description of the problem from the perspective of the engineer, the mathematician or the numerical analyst. A complete modelling tool should incorporate each model as a possible input level. The paper discusses how each model can be represented in a simulation modelling system, and how transformations between models can provide expert guidance to the user through the intermediate models to the final code. Thus, a modelling tool which reduces conceptual distance and guides the search for the best model is proposed.

The remainder of this paper is outlined as follows.
Section 2 discusses the conceptual models between a real-world problem and simulation code, and based on this, an overall system architecture is outlined. A knowledge ontology by which the conceptual models can be realised is described in Section 3. Section 4 discusses inference and model transformation and finally, the current status of a prototype system which implements the above proposals is reported.

2. Solution Perspectives for Numerical Simulation

2.1 User Group Classification

Potentially, numerical simulation may be of interest to three different user groups, numerical analysts, mathematicians and engineers. Each of these groups has different motivations for using simulation, has different expertise levels and has different requirements of the simulation results. Design of any potential modelling environment for numerical simulation therefore, should place emphases on integrating these various perspectives and trying to develop a system that allows exploitation by all user groups.

2.2 User Group Requirements.

The primary concern of the numerical analyst is the development of strong numerical algorithms to deal with difficult numerical properties of the system of equations being analysed. A modelling tool for this type of user should remove the arduous task of coding, while still offering the flexibility to choose or create algorithms and modify important parameters. Mathematicians are interested mainly in the properties of the equations themselves, and may not be familiar with the techniques used to simulate the solution. These users will expect flexibility in specifying mathematical problems. Engineers and scientists are concerned with the modelling of physical problems

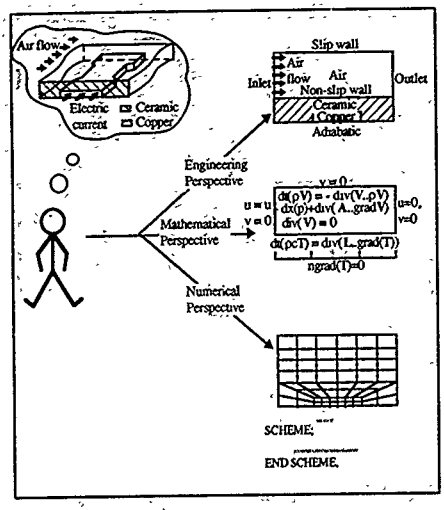


Figure 1 Modelling Perspectives of Real-World Problems

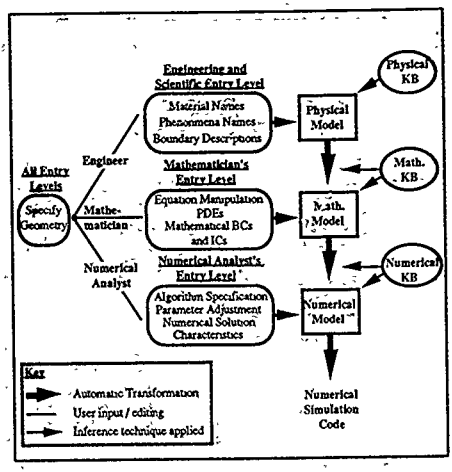


Figure 2 System Architecture

but may not be aware of the underlying mathematical formulation. Thus there are three conceptual levels of understanding, corresponding to the levels of expertise of the different user groups. A real world problem may be modelled from an engineering, mathematical or numerical perspective (see Figure 1) In fact, these perspectives represent the stages through which that problem must be transformed, before it can be simulated on computer.

2.3 A modelling environment for numerical simulation

Figure 2 illustrates a practical proposal of a system architecture which provides a modelling environment suitable for numerical analysts, mathematicians and engineers. The system structure is based on the conceptual levels outlined above. Three representations are allowed, referred to as a *physical model*, a *mathematical model*, and a *numerical model*. If the user is a numerical analyst, the input is a numerical model in the form of a set of numerical algorithms to be solved by a numerical simulation engine. If the user is a mathematician, the input consists of a mathematical representation of a PDE based problem. The system has an incorporated knowledge base (KB) of numerical expertise which automatically transforms this mathematical model representation to a suitable numerical model capable of being solved by the numerical simulation engine. If the user is an engineer, the input consists of a physical model representation which must be brought through a two stage transformation using the expertise of both mathematicians and numerical analysts in two knowledge bases.

3. Knowledge Representation

This section presents a knowledge ontology by which the conceptual models of the previous section are described. The issue is to find an appropriate internal representation of domain knowledge, so that models entered by the user can be checked for consistency and transformations to lower level models can be obtained. A frame-based approach to knowledge representation provides the most natural solution [Tello 1990]. Three knowledge bases, the Physical KB, Mathematical KB and Numerical KB, hold a representation of the problem domain from the perspective of the corresponding model. Conceptual entities which form the basic building blocks of any particular model are stored as frames in each knowledge base. Each entity has a number of attributes associated with it. These may be defined as slot variables within each frame or more generally as functions of slot variables. Moreover, structural information is stored through hierarchical links between these frames.

3.1 Hierarchical Links

Three types of links are found to be of use in the knowledge bases, namely, *taxonomic* links, *component* links and *possible component* links. The latter two are different types of substructure links.

(i) Taxonomic links

Domain concepts fall naturally into categories or classes. For example, while there are many types of PDEs, all PDEs can be expected to share some common features. It makes sense to define a superclass of PDEs and allow all PDEs to inherit common information from this superclass. Taxonomic links, therefore, allow for an efficient storing of information.

(ii) Component links

Component links express the usual component relationships, for example, a region must always consist of a number of boundaries

(iii) Possible Component links

Possible component links express that it makes sense, in a particular model, for a certain entity to be a component of another entity For example, a source term, *may* be a component of a heat equation, but this is not always the case, and depends on the particular problem being modelled. These links play an essential role in inference, by restricting the search space during model transformation.

3.2 The Knowledge Bases

Each knowledge base is now described in more detail

(i) Physical KB

Figure 3 shows part of the hierarchy of the Physical KB Frames represent engineering concepts and terminology There are five main classes, namely, *physical region*, *physical boundary*, *physical boundary condition*, *physical phenomenon* and *geometry* The boundary condition frames correspond to engineering descriptions of boundary conditions, for example, *inflow* or *outflow* in fluid flow

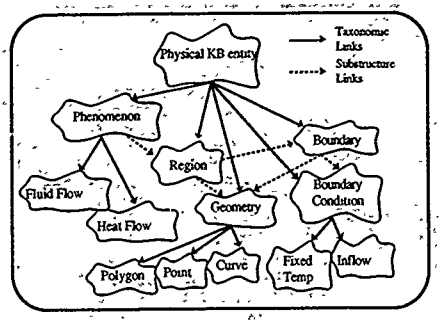


Figure 3 .Physical KB

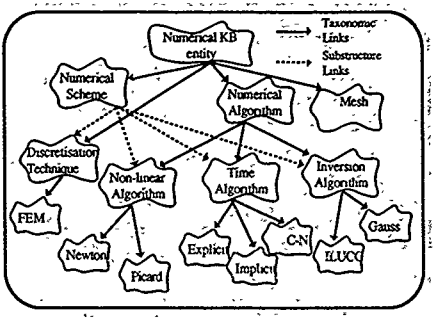


Figure 5 Numerical KB

problems, or *fixed temperature* in heat transfer problems. The *physical region* frame contains a material slot corresponding to the material of which the region is composed. The *physical phenomenon* frame contains a slot for initial conditions

(ii) Mathematical KB

All information corresponding to the representation and structure of PDE problems is contained in this KB. Classes include *PDE group*, *PDE*, *Term* and *Variable*. Many physical phenomena, such as fluid flow or structural analysis, can be described by a particular set of PDEs. These sets are stored as subclasses of *PDE group*. Substructure links, to PDEs, terms and variables, follow in the natural way. It should be noted that the substructure links are *possible component* links. Depending on the particular problem, some, but not necessarily all of these links, will be instantiated in the actual mathematical model, e g. time terms are included only if the problem is transient. Figure 4 shows part of the KB corresponding to the Navier Stokes equation group.

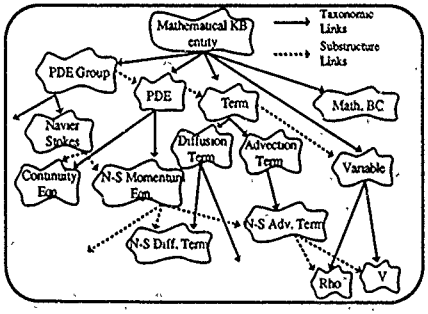


Figure 4 Mathematical KB

(iii) Numerical KB

Frames in the Numerical KB represent the basic numerical algorithms from which a simulation code can be generated. Note that these are not the algorithms themselves, but rather representations which hold essential information about their selection and use. The discretisation technique may be fixed as, say, the Finite Element Method, although, if the simulation engine can cope with different techniques, then the knowledge base should be expanded to include these. Algorithms are organised into classes according to their purpose in the simulation code. Thus, there is a class of algorithms to deal with time-dependence, a class to deal with non-linearity and a class of inversion algorithms which solve the discretised equations. A numerical scheme, that is, a complete numerical method to solve a particular PDE problem, may possibly consist of algorithms from each of these classes (see Figure 5).

To generate a model corresponding to a particular problem, the appropriate frames are chosen from the corresponding KB, and instantiated as components of the model. Slots in the frame must be filled according to problem data. In the next section, the process of transformation from a given model input to other lower level models, is discussed.

4. Model Transformations

4.1 Localised Rule Bases

In order to obtain a lower level model from a high level model, it is necessary to provide bridges between concepts in different knowledge bases. These bridges are realised through localised rule bases bound to each frame definition in each knowledge base. The rule base contains the conditions under which the frame should be instantiated as part of the model. Antecedents of each rule relate to characteristics of the previous model. For example, the rule base bound to the *advection term* in the Mathematical KB may contain the rule

IF *peclet number is not negligible* THEN *select advection term.*

The *peclet number* is a function of attributes of the physical region such as material and dimensions. Functions to calculate such characteristic values are bound to the frame definitions in the knowledge bases.

4.2 Inference Mechanism

The primary goal during model transformation is to find a lower level representation of the given high level model. Substructure links in the knowledge base corresponding to the lower level model, provide subgoals to the primary goal. For example, one subgoal in the physical to mathematical model transformation is to generate a PDE. To achieve this, the further subgoal of generating the PDE's component terms arises. The *possible component* links restrict the space of possibilities for this subgoal. By evaluating the rules in the rule base bound to each possible term, the selected terms are instantiated in the mathematical model and *component* links are established between these and the parent PDE.

4.3 Physical to Mathematical Model transformation

In general there is a one to one correspondence between physical model concepts and mathematical model concepts e g one *PDE group* describes one particular *physical phenomenon*. Generating the mathematical model then follows the inference method outlined above. It is also necessary to link material names from the physical model with numerical values of constants in the mathematical model

A material properties database provides the primary relationship between these descriptive engineering material terms (e.g., copper, air) and the precise numerical property values (e.g., copper conductivity 400 W/m K).

4.4 Mathematical to Numerical Model transformation

In the mathematical to numerical model transformation, decisions are closely inter-related and cannot be made in a simple serial fashion. For example, if the inversion algorithm is always chosen before the time algorithm, and the time algorithm is always chosen before the non-linear algorithm, it may be impossible to decide the best combination of algorithms. To cope with this difficulty, the inference mechanism is allowed to leave a number of possibilities at each decision stage. Thus a solution space of all reasonable algorithm combinations is generated. In this context, the local rules serve to eliminate bad selections, rather than select the best candidate. Each branch of the solution space represents one possible numerical scheme.

An appropriate spatial mesh and possibly a time step must also be derived. One approach is to select the spatial mesh from a fixed family, parameterized by a variable representing the mesh size. The issue of selecting mesh and time step then becomes one of optimization, choosing the best mesh size and time step according to accuracy, stability and efficiency criteria. Each solution branch is rated according to these criteria, but the final choice is left to the user.

4.5 Creation of Numerical Simulation Code

A numerical model contains sufficient information to generate the simulation code. Local code generation procedures are bound to each algorithm frame in the Numerical KB, and a global procedure controls how each part is entered into a code template. These procedures can be tailored depending on the particular simulation code required. In a prototype system, the DEQSOL simulation code is generated. This only requires manipulation of the mathematical representation of the problem, for example, the *EXPLICIT algorithm* frame contains a procedure which replaces the object variable in each differential operator with the variable which holds the old value (e.g. replace div[k grad U] with div[k grad U_0]).

5. Implementation and Current Status

A first prototype of the proposed system has been implemented on a Macintosh using Object Lisp, an object-oriented extension to Common Lisp. The final output of this system is a Finite Element simulation code in the syntax of DEQSOL. Currently only a physical model entry level has been implemented and the user enters the model through a graphical interface.
The knowledge bases are limited to steady state and transient problems in the heat transfer and fluid domains, with several typical algorithms provided. Only two-dimensional geometries, with straight line boundaries are allowed.

6. Conclusions

The complete modelling process involved in transforming a real world problem into a form suitable for numerical simulation, has been analysed and on this basis, three conceptual models (physical mathematical and numerical models), corresponding to the different levels of expertise expected of different types of users, have been identified. A system, based on these conceptual models, in which each type of user can express the problem using familiar terminology, has been proposed. To realise this system, a frame-based knowledge representation approach has been adopted. Domain concepts which form the building blocks of each model have been isolated and represented as frames. For a particular problem, each model is instantiated from a knowledge base, consisting of these frame concepts, stored in a hierarchy. Hierarchical links express the intrinsic structures which any valid model must obey. Transformations between models are achieved through localised rule bases, which form bridges between concepts in different knowledge bases. Transformations may generate more than one possible model, in which case the user makes the final decision. Guidance is offered by rating each possibility according to predicted accuracy as well as efficiency criteria.
The authors believe that the proposed system forms the basis of a totally integrated problem solving environment. However, in order to realise a truly powerful system, the next stage must include intensive knowledge acquisition, in order to develop rules that can deal with the complexities of real world problems.

Acknowledgements

The authors would like to express their appreciation to Dr. J.B Grimson and Professor J.G. Byrne of Trinity College, Dublin for their advice and encouragement. We would also like to thank N Hataoka, Hitachi Dublin Laboratory, for his helpful comments.

References

[Barras *et al.* 1990] Barras, P., Blum, J., Paumier, J.C., Witomski, P. and Rechenmann, F. "EVE. An Object-Centred Knowledge-Based PDE Solver". In *Proceeding of the Second International Conference on Expert Systems for Numerical Computing*, Purdue University, USA, 1990.

[Chelsom *et al* 1990] Chelsom, J., Cornah, D. and Reid, I. "Numerical and Statistical Knowledge-Based Front-ends, Research and Development at NAG". In *Proceeding of the Second International Conference on Expert Systems for Numerical Computing*, Purdue University, USA, 1990.

[Kon'no *et al* 1986] Kon'no, C., Saji, M., Sagawa, N. and Umetani, Y. "Advanced Implicit Solution Function of DEQSOL and its Evaluation". In *Proceedings of IEEE Fall Joint Computer Conference*, pp.1026-1033, Dallas, USA, 1986

[Rice 1985] Rice, J.R. "ELLPACK An Evolving Problem Solving Environment for Software Computing" In *Proc of IFIPTC2, WG 2.5 Working Conference on Problem Solving Environments for Scientific Computing*, pp. 233-245, 1985.

[Russo *et al* 1987] Russo, M., Peskin, R., Kowalski, A. "A prolog-based expert system for modeling with partial differential equations" In *Simulation, Vol 49, No 4*, pp.150-158, 1987.

[Schonauer and Schnepf 1987] Schonauer, W. and Schnepf, E. "Software Considerations for 'Black Box' Solver FIDISOL for Partial Differential Equations". In *ACM Trans. on Maths. Soft, Vol 13, No. 4*, pp. 233-245, 1987.

[Tello 1990] Tello, E.R. "Object Oriented Programming in Artificial Intelligence". Addison-Wesley, New York, 1990.

[Umetani *et al.* 1985] Umetani, Y *et al.* "DEQSOL A numerical Simulation for Vector/Parallel Processors" In *Proc of IFIPTC2, WG 2.5 Working Conference on Problem Solving Environments for Scientific Computing*, pp. 147-164, 1985.

An evaluation tool for CCND language applications

Rajiv Trehan

Information Systems Laboratory
Toshiba Research and Development Center
Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku
Kawasaki-shi, Kanagawa 210, Japan

Abstract

Currently the language and execution models being adopted for the Committed Choice Non-deterministic (CCND) languages are settling down to subsets of the possible models. In general the subsets being adopted are governed by implementation issues rather than application requirements. To give an applications perspective on alternative execution and language models we develop a new evaluation system. The system allows us to consider how an application behaves on alternative language and execution models. Hence providing some applications rationale for the design and implementation alternatives for this class of languages.

1 Introduction

Currently the language and execution models being explored for the Committed Choice Non-Deterministic (CCND) logic languages [10] [17] [5] are settling down to a subset of the possible models. For example, the language executions are being restricted to only investigate clauses sequentially. In general the subsets being considered are governed by implementation issues rather than application requirements. To give an applications perspective on alternative execution and language models we develop a new evaluation system. The system allows us to consider how an application behaves on alternative language and execution models. Hence providing some applications rationale for the design and implementation alternatives for this class of languages.

The execution alternatives considered are for the entire class of CCND languages, rather than any given subset, and represent extremes in the implementation options, for instance the alternatives for scheduling are busy and non-busy waiting. Of course this approach allows us to also compare given subsets of the CCND languages, like flat guards [7], [3]. To compare alternative language and execution models we propose an extended set of evaluation parameters, which rationalise currently accepted parameters for evaluating various sub-classes of the CCND languages.

As with current evaluation systems we attempt to measure the inherent parallelism available in the evaluation of programs. A measure of the inherent parallelism has several uses: it gives a theoretical measure of parallelism against which particular implementations can be gauged; and it provides information for programmers on the relative merits of various programming techniques.

For typical Computer Science application areas, such as matrix multiplication, it is often possible to obtain theoretical measures for the inherent parallelism. However, for AI type problems the parallelism depends on several factors, such as data structures (knowledge representation), inference mechanisms and irregular search spaces. The irregular nature of AI problems makes a theoretical measure of parallelism difficult. Another approach to obtaining measures of inherent parallelism for both regular and irregular problems is to simulate the computation on an infinite processor model. The simulated processor utilisation gives a measure of the inherent parallelism. It is this second approach we adopt in this work. To obtain a measure of the inherent parallelism available in program execution we adopt a breadth-first execution model assuming an unlimited number of processors.

We first consider current models of execution and the parameters collected during program execution and discuss their limitations in measuring inherent parallelism. The limitations highlighted are then addressed in two respects. Firstly, we develop a new system which allows us to more accurately measure the inherent parallelism in the execution of a program. Secondly, we consider possible alternatives open to language implementors, for instance different scheduling policies. This provides the basis for a set of parameters which can be used to indicate the relative merits of the alternatives. The measurement of the proposed parameters is carried out for one CCND language, Parlog [5]. We then present a profiling tool, developed to graphically display profiles of the various proposed parameters over time (cycles). Evaluations using this tool are given for a small set of simple example programs and the results analysed. Finally we consider limitations of the system and areas of future work.

2 Current measurements and their limitations

The evaluation/comparison of application level programs tends to be based on course grained metrics/parameters, like logical inferences for Prolog [20]. For the CCND languages three parameters are usually quoted when comparing applications, namely cycles, suspensions and reductions [10] [11] [8] In general these parameters appear to be collected using interpreters on top of Prolog [10] [5] [12], although they could be collected using compilers [5] [18] [9] or abstract machine emulators [1] [2] [6] [19].

The evaluation/comparison also tends to be based on the inherent parallelism available at the applications level [10] [11] [8]. This is obtained by using a breadth-first evaluation model assuming an infinite number of processes. There are two main limitations with the current parameters. Firstly, the actual evaluation models employed make several approximations to a full breadth-first evaluation

- the AND-parallel goals are represented as a list of goals to be evaluated;
- as each goal is reduced, the resulting body goals are added to the goal list and any appropriate bindings are made;
- variable bindings are produced in the order that the goals are evaluated;
- guarded goals are evaluated as a single reduction which incur no cycle overheads;
- the interpreters make no distinction between suspension and failure of guarded goals;
- the interpreters model OR-parallelism by backtracking through alternative clauses; and
- the first textual clause in a predicate whose guarded goals succeed is committed to.

The second is that the inherent parallelism depends on choices made about scheduling, guard termination and suspension mechanism, however, the effect of these alternatives are not considered or reflected in the current evaluation parameters[1].

In the following sub-section we consider how the approximations to a breadth-first model may introduce erroneous data into our evaluation parameters. We then develop an improved system which addresses these execution limitation. Finally, we propose and collect an extended set of evaluation parameters which reflect execution alternatives.

2.1 Cycles

The cycles parameter attempts to measure the depth of the breadth-first execution tree. A cycle corresponds to reducing all the goals in the system once in parallel, that is assuming an infinite number of processors.

The evaluations of guarded goals are carried out by a call to the top-level of the interpreter and for simplicity is assumed to take *zero* cycles to evaluate (the guard evaluation is assumed to be part of the commitment). Hence, the cycle count can only claim to measure the depth of the evaluation tree when evaluating flat code [7], [3]. Moreover, in the case of deep guards, any goals suspended awaiting the evaluation of a guard will only suspend for one cycle and not the number of cycles it takes for the deep guard to be evaluated. This distorts the breadth-first evaluation tree, reducing the cycle count.

Another limitation is that the goals are maintained as a list which is processed in a left-to-right order, any bindings made in the evaluation of goals taking place immediately. So, these bindings will be available to any remaining goals in the goal list. This will allow goals that require these bindings to reduce in the current cycle. Hence the evaluation is dependent on goal order.

[1] Of course there are also several other limitations, like the use of an infinite processor model. However, this model provides a yard stick against which particular processor implementations can be judged.

2.2 Reductions

The reductions parameter attempts to measure the commitments performed by the system in solving a query, which indicates the number of parallel goal evaluations that can take place.

The first limitation is that current interpreters try evaluating the alternatives for a given predicate top-down, committing to the first clause whose guarded goals succeed. Therefore reductions can only be counted for the clauses that have been attempted. Hence the reduction count depends on the clause order.

Another problem occurs if a goal evaluation fails, it is re-scheduled and may be re-attempted (if the computation goes on for further cycles before deadlocking). The re-evaluation of failed goals may introduce erroneous statistics into the reduction count.

Finally, the evaluation of system calls, supported by calls to the underlying Prolog, are not accounted for. These do contribute to the overall work done in evaluating programs. By ignoring their contribution the comparison of programming techniques which make use of system primitives can be misleading.

2.3 Suspensions

The suspensions parameter attempts to count the number of suspended evaluations in the evaluation of a query.

The number of suspensions depends on when suspended evaluations are re-scheduled. When evaluations suspend in the existing interpreters they are immediately re-scheduled for evaluation; this is known as busy waiting. However, a non-busy waiting strategy could have been used[2]. Using an ideal non-busy waiting strategy each evaluation will only suspended once.

Another point to note is current interpreters process the goals from left-to-right, generating bindings as the goals are processed. These bindings may allow goals to reduce in the current cycle which would suspend if the goals were evaluated in a different order.

Finally, as no distinction is made between failed and suspended goals, erroneous statistics may be introduced into the suspension count.

3 An improved execution system

3.1 Requirements

Many of the limitations and inaccuracies of the statistics obtained using current implementations are due to features of the execution model employed. The collection of more meaningful statistics requires the development of an improved implementation. Such an implementation would have to exhibit the following features:

- It must distinguish between suspension and failure of a goal evaluation.

[2] The suspended evaluations are hooked (or tagged) to the variables that were required when they suspended, when sufficient variables become instantiated they are re-scheduled.

- It must more accurately measure the depth of the evaluation tree: the depth of the evaluation tree should not be dependent on goal or clause order; and the depth of the evaluation tree must account for the use of deep guards.
- It should realise AND-parallelism: each of the goals in the conjunction should appear to be evaluated in parallel; each AND-parallel goal should be reduced once in each cycle. The reduction may take place in the guarded evaluation in the case of deep guards; and the simulated reduction of each goal in a given cycle should be independent of the actual order in which the goals are processed.
- It should realise OR-parallelism: each of the clauses that a goal could use to reduce should appear to be explored in parallel; in a parallel evaluation a goal should commit to the first clause whose guard successfully terminates; and the evaluation of a goal suspends if no committable clause exists and at least one clause evaluation suspends.

3.2 Idealisations

We now consider idealisations made in the design of our improved execution model.

3.2.1 Guard evaluation idealisations

We considered two alternative models for incorporating the effect of the guard evaluation into the overall cycle based evaluation[3]. The first assumes that in a cycle, a goal can be head unified with a clause and the guarded goal evaluation instigated. This model assumes that guarded system goals (flat guards) will evaluate in 1 cycle. The second assumes that in a cycle, a goal can be head unified with a clause and either the guarded evaluation instigated or a system guard evaluated. This model assumes that system goals incur no cycle costs. We adopt the second model as this model has a similar notion of depth to the previous implementations when executing flat guarded programs.

3.2.2 AND-parallel idealisations

A fully accurate model would be able to determine exactly when a goal makes a binding, how long it would take for this binding to reach another goal and whether this would be in time for the goal to use it. Clearly the inherent parallelism should not be dependent on goal order. We make the assumption that, in a cycle, a goal can only use bindings available to it at the start of the cycle. Such a model may not display all the parallelism that could be achieved in a given implementation, but at least it gives a measure which is not dependent on how the goals are ordered.

[3]It should be noted that most cycle based models for obtaining a measure of the depth of the evaluation tree will be prone to giving distorted results, in that they will tend to associate fixed costs with the various components of the evaluation, like head unification, system call evaluation and commitment. Although an elaborate model of cost could be developed, these costs would tend to be implementation dependent. Moreover, such a model is unlikely to give a radically different view of the general features of the evaluation compared with a fixed cost model.

3.2.3 OR-parallel idealisations

The modelling of inherent OR-parallelism requires the evaluation to commit to the clause whose first guard successfully terminates. In our system the duration of a guard evaluation is approximated by the depth of its evaluation tree. So the evaluation should commit to the guard with the shallowest evaluation tree.

3.2.4 System call idealisations

System calls contribute to the overall work done in the evaluation of a goal. However, the complexity of each system call may vary. This makes the matter of system calls difficult to address, as it will be implementation dependent and there is no agreement on the nature of such calls. As a general principle we count system calls as reductions, as they contribute to the overall work done (although our evaluation system is easily adapted to ignore all or some of the system calls, see section 5). In suspending the evaluation of a system call we assume it behaves like a goal with one clause.

3.3 Development/Realisation

We could implement the required improved execution model by an abstract machine emulated in 'C'. The abstract machine could then be instrumented to collect, dump, dynamic information like the cost of various operations and the reference characteristics of our abstract machine and implementation, in the same vein as [13]. However, this approach was not feasible for this work because there was, and still is, no representative abstract machine for the entire class of CCND languages available for instrumentation.

Alternatively, we could implement an improved interpreter which could either dynamically collect statistics like the previous evaluation interpreters, or dump information about the program evaluation as in a 'C' emulator. We have chosen to implement an interpreter which will dump data about the program execution. The dump data is rich enough to allow flexible analysis assuming alternative models of execution. Hence we are able to support an extended set of evaluation parameters.

Using an interpreter allows us to collect coarse grained information like number of commitments or size of suspension queues. These coarse parameters are similar to the coarse grained parameters, like logical inferences [20], collected for sequential logic programming languages (Prolog) and to the currently accepted evaluation parameters used for the CCND languages [10] [11] [8]. However, an interpreter does not easily allow us to measure fine detail, like [13].

The incremental design and development of the improved interpreter can be found in [16]. The features of this interpreter are as follows:

- both AND and OR-parallelism are modelled;
- each of the guarded goals for a given predicate is tried and relevant statistics collected;
- the statistics from the evaluation of the guarded goal are used to pick the solution path (currently this is the shallowest successful guard, *i.e.* the first guard that would have succeeded in a breadth-first execution);
- the goals that form the goal list undergo one reduction in a cycle (any bindings made as the goal list is processed occur only when all the goals have been attempted);
- the evaluation of a system goal which makes a call to the underlying Prolog system is counted as one reduction;
- bindings made using calls to the underlying Prolog system are made only when all the goals have been attempted;
- the interpreter makes a distinction between suspension and failure;
- although guard evaluations are carried out to completion in one go, the commitment of a goal to a given clause is prevented for the number of cycles the guard took to evaluate; and
- the suspension of all the guarded goals causes suspension of the goal being evaluated.

4 New evaluation parameters

Apart from having inaccuracies in measuring the inherent parallel behaviour of programs introduced through limitations in the execution model, the parameters proposed by Shapiro (cycles, suspensions and reductions) do not give any indication of the effects that alternative implementation models may have had.

In this section we consider the extremes of the current implementation models. Our new parameters aim to reflect the effect of adopting different alternative execution models and so give us a better understanding of our applications.

4.1 Basis for new parameters

4.1.1 Pruning OR-branches

The parallel evaluation of a goal invokes several guard evaluations, one for each clause that the goal successfully head unifies with. The evaluation commits to the first clause whose guarded system successfully terminates. On commitment, the other guarded systems invoked by the goal evaluation can be terminated or ignored. Terminating the alternative clauses (pruning) requires the system to stop the computation being carried out in the alternative branches. This may prevent these branches carrying out needless computation. Ignoring the other alternative clauses (non-pruning) when a goal commits, requires the system to disregard any commitment requests from the other alternatives should their guarded systems also terminate successfully. This assumes that guard evaluations terminate and certainly do not diverge. This may save some computation (in sending a terminate message to the other guard evaluation) if the guards are balanced or in cases where only one clause can be committed to.

4.1.2 Suspension mechanisms

A goal evaluation suspends if there is no committable clause and at least one of the guard evaluations or head unifications suspends. Suspending the evaluation can be achieved in several ways, the two extremes being goal suspension and clause suspension. Goal suspension involves suspending the parent goal of a computation when each of the clause evaluations suspend.

Alternatively each of the clauses (guarded computations and head unifications) could be suspended, which is known as clause suspension. As there may be recursive guard evaluations invoked, clause suspension may result in a tree of suspended evaluations, representing the guard call structure. The trade-off between these two extremes is basically a space-time consideration. Suspending a goal requires less space than suspending the evaluation of each of the clauses. However, if some computation is performed in the evaluation of the guard before suspension then this computation will be lost, and repeated, if the goal is suspended.

4.1.3 Scheduling policy

Another choice is how and when suspended evaluations are re-scheduled. When an evaluation suspends it could be tagged to the variables which are required and unbound, and re-scheduled when they become bound, this is known as non-busy waiting. It should be noted that some predicates, like `merge/3`, only suspend on one variable whereas others, like `equals/2`, require both arguments to be bound. The other extreme would be to immediately re-schedule the suspended evaluation, known as busy waiting.

Employing a non-busy waiting suspension mechanism is appropriate if suspended evaluations remain suspended for several cycles, for example in generating prime numbers by sifting [5] most of the filter processes will be suspended most of the time. Employing a busy waiting suspension mechanism is appropriate if suspended goals are only likely to be suspended for a short period (see [14], [15]), as with Layered Streams [8].

4.2 Proposed profiling parameters

Our new profiling parameters aim to reflect the effect of various alternative execution options We propose collecting suspensions and reductions using combinations of the following alternatives. OR-branches may be pruned or non-pruned, suspended evaluations may be goals or clauses and scheduling may be busy or non-busy The new parameters are suspensions and reductions using

- busy waiting, non-pruning and goal suspension,
- busy waiting, non-pruning and clause suspension,
- busy waiting, pruning and goal suspension;

- busy waiting, pruning and clause suspension;
- non-busy waiting, non-pruning and goal suspension;
- non-busy waiting, non-pruning and clause suspension;
- non-busy waiting, pruning and goal suspension; and
- non-busy waiting, pruning and clause suspension.

A full description of all 8 models of execution and addition parameters we also consider, th depth of the evaluation and the minimum reductions, is given in [14].

5 Profile tool

As mentioned earlier, see section 3.3, our new interpreter provides information about the execution by creating a dump-data file. This dump file contains tokens which allow us to build a parallel picture of the execution under a range of alternative models of execution. The tokens in this dump file are used by our post analysis to extract the parameters we put forward. The main features of this post analysis are:

- The profiler maintains an aggregate of each of the proposed profile parameters per cycle. This provides us with a break down of the various profiling parameters which can be used to give a dynamic picture of the execution. Moreover this mechanism also provides the means by which we are able to collect pruned/non-pruned, busy/non-busy and goal/clause data.
- Profiling the guard data produces a cycle by cycle aggregate of each of the parameters as collected in the guard evaluation.
- On completing the profiling of some guard data the next token indicates that the parent goal either commits, suspends or fails.
  - If the goal commits, the clause number and depth of the commitment are also returned. This provides information which is used to *prune* those profiling parameters which adopt a pruned execution model. The *pruned* guard profile is then combined (spliced) into its parent's profile data.
  - If the goal evaluation suspends for the first time the suspension parameters of the guarded goal evaluations are spliced into the parent's profile. An additional suspension is also added to all the goal suspension parameters, at the depth at which the guard suspended.
  - If the goal evaluation re-suspends, the busy suspension parameters of the guarded goal evaluations are spliced into the parent's profile. An additional suspension is also added to the busy waiting goal suspension parameters, at the depth at which the guarded evaluation suspended.

We have also implemented a profile tool which executes under SUNVIEW™. This tool allows us to see any, or several, profiling parameters in graphical form. The tool also provides information on the totals of the various parameters. 32 The dump file could be used to collect several other parameters. For example, we count system calls as reductions, which we use as a measure of parallelism. However, the dump file contains different tokens for commitments and system call evaluations and so different measures for the parallelism could be obtained.

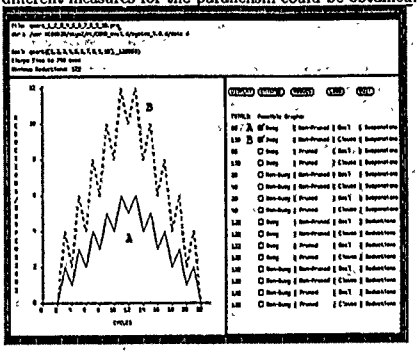


Figure 1: Example of the Profiletool for Quick-sorting

Figure 1 contains an example screendump of our tool. The tool shows plots of the number of reductions and suspensions (y-axis) in each cycle (x-axis). Such plots can be given for any combination of suspension mechanism, scheduling strategy and pruning option. The options selected are indicated by the toggle buttons on the right hand side. The tool also presents information on the total number of reductions and suspensions using a given execution model, next to the toggle buttons for each option. Finally, the tool also contains some more general information, such as: the goal that was evaluated; the elapsed time of the evaluation, and the minimum number of reductions required to evaluate the goal, assuming the existence of an oracle to pick the correct clause.

6 Example executions and measurements

In this section we consider the behaviour of two simple example programs, "quick sort" and "prime number generation". These programs are simple enough to be considered theoretically and so we use them to highlight the improvement and use of our profile tool; The evaluation of other programs including a collection of AI programs can be found in [14] and [15].

6.1 Quick-sort

The quick-sort example highlights the various differences in the various suspension parameters and how reg-

```
mode quicksort(?,^).
quicksort(Unsorted,Sorted):-
    qsort(Unsorted,Sorted-[]).

mode qsort(?,^).
qsort([X|Unsorted],Sorted-Rest):-
    partition(Unsorted,X,Smaller,Larger),
    qsort(Smaller,Sorted-[X|SortedTemp]),
    qsort(Larger,SortedTemp-Rest).
qsort([],Sorted-Rest):- Sorted = Rest.

mode partition(?,?,^,^).
partition([X|Xs],A,Smaller,[X|Larger]):-
    A < X :
    partition(Xs,A,Smaller,Larger).
partition([X|Xs],A,[X|Smaller],Larger):-
    A >= X :
    partition(Xs,A,Smaller,Larger).
partition([],_,[],[]).
```

Figure 2: Quick-sort program in Parlog

ular and irregular queries result in differing dynamic features of the computation. Firstly, we consider how this program behaves if the list to be sorted is already ordered and then if the input list is unordered.

6.1.1 Quick-sort of an ordered list

Consider the program in *Figure 2* with the following query

```
quicksort([1,2,3,4,5,6,7,8,9,10],L)
```

The regular nature of the data for this query allows us to reason about its evaluation. Basically, `quicksort([1,2,3,4,5,6,7,8,9,10],L)` will be reduced to the initial `qsort/2` goal. This goal is then reduced to a `partition/4` and two new `qsort/2` goals. The `partition/4` goal will partition the input list (based on the current first element; the *pivot*) into two output lists. One output list contains elements greater than the *pivot*, the other elements less than the *pivot*. As the input list is already ordered the `partition/4` goal will only add elements to one of the output lists. The two `qsort` processes will initially suspend awaiting the output lists from the `partition/4` to be generated. In the following cycle one of the `qsort` goals will be able to reduce, as the `partition/4` process constructs the output lists. The reduction of this `qsort` goal will again result in a `partition/4` process and two `qsort/2` processes. The other `qsort/2` process remains suspended until the entire list has been partitioned, *i.e.* until the `partition/4` processes complete.

These processes will behave as before. the `partition/4` process will only add elements to one of its output lists, the two `qsort` processes will initially suspend, one of which will be able to reduce in the following cycle. The partition processes will be spawned in cycles 2,4,6,8,10,12,14,16,18,20 respectively. The duration of the processes will be 9,8,7,6,5,4,3,2,1 cycles respectively. Hence, these processes will terminate in cycles 11,12,13,14,15,16,17,18,19,20,21. After cycle 11 there will be one less suspended process each cycle (a `qsort/2` process) until cycle 21.

This effect has to be combined with the spawning pattern of the `qsort/2` goals, *i.e.* initially 2 suspensions, of which 1 reduces in the next cycle. The overall goal suspension pattern is, initially in every other cycle there will be two new suspensions, one of which is able to reduce in the following cycle. After cycle 11 the pattern will invert. In every other cycle there will be one new suspension followed by two of the suspended goals being re-scheduled and reducing.

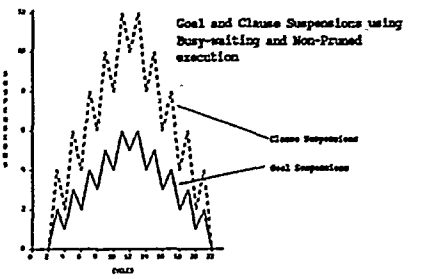


Figure 3. Quick-sorting an ordered list (goal and clause suspensions)

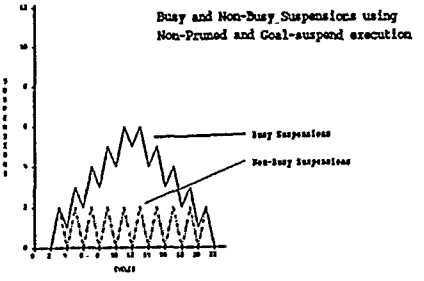


Figure 4. Quick-sorting an ordered list (busy and non-busy suspensions)

Figure 3 gives profiles for goal and clause suspensions using busy waiting and non-pruning Using busy waiting gives us a measure of the total number of processes suspended. Note that the goal suspension profile (the

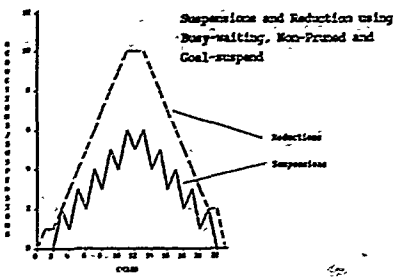


Figure 5: Quick-sorting an ordered list (reductions and suspensions)

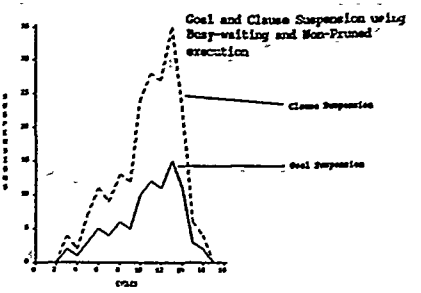


Figure 6: Quick-sorting an unordered list (goal and clause suspensions)

lower graph) is as predicted, that is the total number of suspended processes will initially increase in a step wise manner (steps of +2, -1) and from cycle 11 onwards will reduce in a step wise manner (steps of +1, -2).

Moreover comparing the goal and clause suspension profiles indicates the number of clauses that each suspended evaluation could be reduced by in the dynamic program [4]. This gives a ratio of exactly 2 clause suspensions for every goal suspension. This is also confirmed by our analysis, in that the only processes to be suspended are `qsort/2`, which could be evaluated via two clauses.

Figure 4 gives profiles for goal suspension using busy and non-busy scheduling strategies. Busy waiting gives a measure of the total number of suspended processes while non-busy gives a measure of the new suspended processes in each cycle. The profiles fit the analysis of this execution. In every other cycle there will be two new suspended goals, one of which will reduce in the next cycle.

Finally we give a profile of the reductions in *Figure 5*: The number of reductions increases by 1 each cycle, as new `partition/4` processes are spawned and reduced. After cycle 11 the `partition/4` goals begin to succeed (terminate) and the processes begin to collapse. At the peak there will be 10 `partition/4` and 1 `qsort` process reducing in parallel.

6.1.2 Quick-sort of an unordered list

We now turn our attention to the behaviour of quick-sort on an unordered list. Consider the the following query:

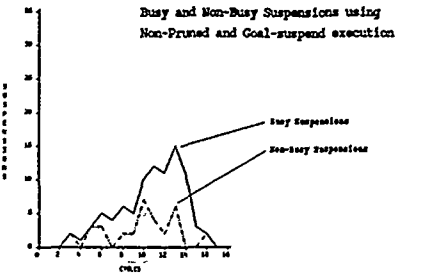


Figure 7: Quick-sorting an unordered list (busy and non-busy suspensions)

```
quicksort([4,6,2,9,5,5,1,10,3,7],L)
```

The irregular nature of the data for this query makes reasoning about its evaluation difficult. However some global features can be predicted, namely:

- The unordered query will result in the `partition/4` process adding elements to both output lists. This will result in both the `qsort/2` processes reducing before the `partition/4` processes have terminated. Compared with the ordered list example this should show an increase in the average number of reductions and reduce the total length of the computation.
- As the `partition/4` processes add elements to both output lists the `qsort/2` processes reduce to three suspended processes, *i.e.* the newly spawned `partition/4` goal may suspend because no further elements have been added to its input list. This will be indicated by the ratio between goal and clause suspensions increasing, as the `partition/4` processes

[4] There is a difference betwe[illegible] counting the nu[illegible]er of clauses for each predicate statically, and the dynamic nature of the program, as some predicates may be used more often than others, hence weighting the results. Of course comparing suspensions for goal and clause suspension mechanisms only provides the dynamic information for suspended evaluations and not the whole evaluation. This comparison still provides useful information about the space-time considerations for the suspension mechanism Suspending the clauses may save head unifications and possibly some reductions (for deep guard examples) but requires more space in that there may be several clause states to suspend rather than a single goal state.

can be evaluated via 3 clauses, whereas the qsort/2 processes can be evaluated by 2 clauses.

- In both the ordered and unordered list examples 10 partition/4 processes are spawned (one for each element of the input list). In the ordered example each partition/4 process will partition the remainder of the input list, *i.e.* the partition process with a *pivot* of 3 will have to partition the remainder of the input list, namely [4,5,6,7,8,9,10]. For the unordered example each partition/4 process will only have to partition a subset of the remaining input list because the remaining input list will be partitioned into two lists. Hence, there will be less reductions performed in the sorting of the unordered list example.

We now compare the data collected using our new profiling system with the theoretical evaluation given above. *Figure 6* gives profiles for goal and clause suspensions using busy waiting and non-pruning. The first point to note is that the ratio of goal to clause suspensions changes from 1:2 for the ordered example, to 90:205 for the unordered example. From this we can conclude that some partition/4 processes suspend. Furthermore, we can see that the overall length of the computation has been reduced from 23 cycles (for qsorting an ordered list) to 18 cycles for this example.

Figure 7 gives profiles for goal suspensions using busy and non-busy scheduling strategies. Firstly, we can see that the duration of suspended processes is more complex to predict. Secondly, we see that the ratio of busy to non-busy suspensions is 90:33 for this query, whereas it was 65:20 for the ordered list example. This indicates that the suspended goals remain suspended for less time in the unordered example, which is intuitively the case.

Finally, comparing the total reductions performed for the ordered list (122 reductions) and the unordered list (72 reductions) we see that, as predicted, there is a marked decrease in the required number of reductions.

6.2 Prime number generation by sifting

The prime number example illustrates how our model for AND-parallelism gives a more realistic indication of the depth of the computation. The program (see *Figure 8*) used generates prime numbers by sifting a stream of integers [17]. As each prime number is produced it results in a filter process being spawned; each filter process removes multiples of itself from the remainder of the stream. So the algorithm involves generating a pipeline of filter processes one for each integer that is unfiltered (new prime) by the previous set of filter processes. We consider the generation of *primes up to 50* and *primes up to 500*.

This algorithm gives a good indication of how the execution model affects the collection of meaningful statistics The technique involves generating a stream of integers, say *fifty*, these integers being generated in *fifty* cycles. This stream of integers then under goes a sifting stage, this will require further cycles. For example, consider the number 47. This will be generated in the *forty-seventh* cycle. It will then be processed by filters representing the following prime numbers. 2,3,5,7,11,13,17,19,23,29,31,37,41,43 This takes at least *fourteen* cycles, one for each filter process.

```
primes :-
    integers(2,I), sift(I,J).

mode integers(?,^ ).
integers(N,[N|I]) :-
    N1 is N+1,integers(N1,I).

mode sift(?,^ ).
sift([],[]).
sift([P|I],[P|R1]) :-
    filter(I,P,R), sift(R,R1).

mode filter(?,?,^ ).
filter([],_,[]).
filter([N|I],P,R) :-
    0 =:= N mod P
    :
    filter(I,P,R).
filter([N|I],P,[N|R]) :-
    0 =\= N mod P
    :
    filter(I,P,R ).
```

Figure 8: Prime number generation by sifting

Previous statistics give a cycle count of only *fifty*, this is because the goals in the process queue are evaluated in a left-to-right fashion. Hence, an integer that is produced in a given cycle is able to propagate through the filter processes in the same cycle (the filter processes in the queue are set-up in a left to right fashion). Our system gives 67 cycles to produce the first 50 prime numbers. *Fifty* of these cycles can be attributed to producing the 50 integers, another *fifteen* can be attributed to propagating the last integer through the *fifteen* filter processes and the remaining *two* are due to spawning the first filter process and terminating the output stream. The effect of accounting for the propagation of the integers through the filter processes results in the number of suspended goals being higher.

7 Limitations of the new measurements

The limitations of our new system can be classified in two ways. Those associated with modelling the execution and the collection of our proposed parameters, and those associated with information we do not or cannot collect

- We adopt a fixed cost model, in terms of cycles. However, the cost of the given operation may depend on

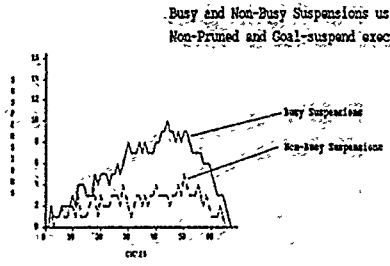


Figure 9: Prime numbers up to 50 (busy and non-busy suspensions)

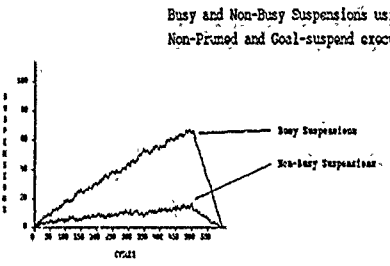


Figure 10. Prime numbers up to 500 (busy and non-busy suspensions)

several factors, such as its complexity. It would be better to adopt a functional cost model, where the cost of an operation is calculated based on its complexity. Such a model would however require the costs of the various operations to be accurately quantified. However such a cost model would be difficult to construct without reference to an actual implementation

- We assume that, in a cycle, a goal can only use bindings available to it at the start of the cycle. While this is an improvement over the current interpreters, in modelling the inherent parallelism. Our model may not display all the parallelism that could be achieved in a given implementation, but at least it gives a measure which is not dependent on how the goals are ordered.

- The new interpreter (like previous interpreters) uses goal suspension. Unlike previous interpreters our suspension record contains information about the duration of the guard evaluation before the evaluation suspended. This information is used to indicate if the process would have been evaluated using clause suspension. Whilst this is an improvement over the previous system our new parameters may be in error for certain classes of program [14].

- The system does not consider possible compile time optimisations, such as clause indexing. For example, the four clauses for a merge goal can be indexed via two switch statements. Of course, this leads to the results for clause suspensions not being directly applicable to compiler implementations using this such indexing. However, it should be noted that the comparison of application programs is still possible, as the program expects four suspended clauses to result in four clause suspensions.

8 Conclusions

Currently the language and execution models being adopted for the Committed Choice Non-deterministic (CCND) languages are settling down to subsets of the possible models. The resulting languages appear to have flat guards and adopt non-busy waiting scheduling policy. There is still some discussion on whether the languages should be safe or unsafe. In general the subsets being adopted are governed by implementation issues rather than application requirements. This work aims to place some applications rationale for the design of language features and their usage. Considering applications is important because the classes of applications that language implementors aim to support and the models of execution that they provide may not be those required by applications programmers.

In this article we present an evaluation system for comparing applications on alternative language and execution models. The evaluation system addresses two classes of limitations with previous systems: The first is that previous systems claim to execute the object code (CCND program) breadth-first (hence allowing the inherent parallel features to be measured), however, the actual evaluation models used make several approximations. The second is that the parameters quoted in the evaluations of a program do not reflect possible alternatives open to language implementors, like scheduling policy. This results in misleading and distorted measurements.

The new evaluation system comprises two stages: an AND/OR-interpreter, which evaluates the program breadth-first producing a dump file, and an analyser program which reconstructs a parallel view of the program execution for different execution models. The statistics obtained are more accurate in two respects. The first is in the modelling of a parallel AND/OR execution, which allows us to measure the inherent parallel features of our algorithm. The second is in observing the affect of alternative models of execution. pruned or non-pruned

guard evaluations; busy or non-busy waiting; and goal or clause suspension.

Finally, we highlight the use and improvement of our system by considering the evaluation of two simple programs. The nature of these programs allows us to consider the theoretical behaviour of their execution which we compare with that predicted by our system. However, this system was built to allow us to compare AI applications, a detailed study of three classes of AI application using the system is given in [14].

Acknowledgements

Firstly, I would like to thank my supervisors Dr. Paul Wilk and Dr. Chris Mellish for all their help and advice. Secondly, I would like to thank Dr. Paul Brna and Dr. Steve Gregory for comments about [14], on which this article is based. Finally, I wish to thank the many people who have read and commented on this work: Andy Bowles, Jim Crammond, Tim Duncan, Julian Gosnell, Ross Overbeek, Hiroshi Sakai, Robert Scott and David Warren.

The work reported in this article was carried out while the author was at the Department of Artificial Intelligence, University of Edinburgh and supported by the Science and Engineering Research Council under a Research Studentship.

References

[1] J. Crammond. *Implementation of Committed Choice Logic Languages on Shared Memory Multiprocessors.* PhD thesis, Department of Computer Science, Heriot-Watt, Edinburgh, May 1988.

[2] I. Foster, S. Gregory, G. Ringwood, and K. Satoh. A Sequential Implementation of Parlog. In E. Shapiro, editor, *Third International Conference on Logic Programming*, pages 149–156, London, 1986. Springer-Verlag.

[3] I. T. Foster and S. Taylor. Flat Parlog: a basis for comparison. *International Journal of Parallel Programming*, 16(2):87–125, 1988.

[4] K. Fuchi and K. Furukawa. The role of logic programming in the Fifth Generation Computer Project. In *Third International Logic Programming Conference, London, United Kingdom*, pages 1–24, 1986. Keynote Address.

[5] S. Gregory. *Parallel Logic Programming in Parlog.* Addison-Wesley, 1987.

[6] Y. Kimura and T. Chikayama. An Abstract KL1 Machine and its Instruction Set. In *1987 Symposium on Logic Programming*, pages 468–477, San Francisco, California, 1987. Computer Society Press.

[7] C. Mierowsky, S. Taylor, E. Shapiro, J. Levy, and M. Safra. The Design and Implementation of Flat Concurrent Prolog. Technical Report CS85-09, Weizmann Institute of Science, Rehovot, Israel, 1985.

[8] A. Okumura and Y. Matsumoto. Parallel Programming with Layered Streams. In *Fourth Symposium on Logic Programming*, San Francisco, 1987.

[9] V. A. Saraswat. Compiling cp(|,|,&) on top of prolog. Technical Report CMU-CS-87-174, Carnegie Mellon, October 1987.

[10] E. Shapiro. A Subset of Concurrent Prolog and Its Interpreter. In E. Shapiro, editor, *Concurrent Prolog: Collected Papers*, chapter 2, pages 27–84. MIT Press, 1987.

[11] L. Sterling and M. Codish. PRESSing for Parallelism: A Prolog Program Made Concurrent. In E. Shapiro, editor, *Concurrent Prolog: Collected Papers*, chapter 31, pages 304–350. MIT Press, 1987. Volume 2.

[12] J. Tanaka, K. Ueda, T. Miyazaki, A. Takeuchi, Y. Matsumoto, and K. Furukawa. Guarded Horn Clauses and Experiences with Parallel Logic Programming. Technical Report TR-163, Institute For New Generation Computer Technology, Tokyo, 1986.

[13] E. Tick. *Memory Preformance of Prolog Architectures.* Kluwer Academic Publishers, Norwell, MA02061, 1987.

[14] R. Trehan. *An investigation of design and execution alternatives for the Committed Choice Non-Deterministic Logic language.* PhD thesis, Department of Artificial Intelligence, University of Edinburgh, Scotland, 1989.

[15] R. Trehan. The Behaviour of Search Algorithms in a Committed Choice Framewor. In *The 1990 Logic Programming Conference*, pages 143–159, Tokyo, Japan, July 1990. Institute For New Generation Computer Technology.

[16] R. Trehan. An evaluation system/tool for CCND language and execution models. *New Generation Computing*, 1991. To appear.

[17] K. Ueda. Guarded Horn Clauses: A Parallel Logic Programming Language with the Concept of a Guard. In Fuchi K and Nivat M, editors, *Programming of Future Generation Computers*, pages 441–456. North-Holland, 1988.

[18] K. Ueda and T. Chikayama. Concurrent Prolog Compiler on Top of Prolog. In *Symposium on Logic Programming*, pages 119–126. IEEE Computer Society, 1985. Also: New Generation Computing, Vol. 2, No. 4, pp 361-369.

[19] D.H.D. Warren. The SRI-Model for OR-parallel Execution of Prolog-Abstract Design and Implementation. In *Proceedings of the IEEE 4th Symposium on Logic Programming*, pages 92–101, San Fransisco, 1987.

[20] P. F. Wilk. Prolog Benchmarking. Research Paper 111, Department of Artificial Intelligence, University of Edinburgh, 1983.

A SIMPLIFIED APPROACH FOR DELAMINATION OF COMPOSITE LAMINATES

L. DAUDEVILLE, P.LADEVEZE
Laboratoire de Mécanique et Technologie / E.N.S.Cachan / Université Paris 6
61, Av du Président Wilson, 94235 Cachan Cedex, FRANCE

Abstract - A new numerical method for delamination of composite laminates based upon a Damage Mechanics approach is proposed. Degradation of the interface between two adjacent layers is introduced. This approach allows an accurate prediction of both initiation and propagation of delamination with a single model. A program has been realized to treat delamination in the vicinity of weakly curved edges with low numerical cost. First results are presented.

I.INTRODUCTION

In the design of carbon-epoxy laminates such as T300-914 the mode of damage and rupture can be split into two classes. The first class is due to fiber rupture, matrix and fiber-matrix interface degradations. The modelling of these degradation phenomena at the single layer scale can be described using Damage Mechanics (Ladevèze,1991), (Allen,1987), (Taljera,1985). The second class of rupture is delamination. This phenomenon which consists in the debonding of layers, appears in the vicinity of the edges of structures.

In the analysis of this phenomenon, distinction is made between a initiation phase and a propagation phase of an existing delaminated area. At present the simulation of both phases relies on the hypothesis of a perfect bonding between elastic layers. The analysis of initiation is quite qualitative, in view of the normal stresses obtained from an elastic computation a more or less important tendency is assigned to different stacking sequences. Empirical criteria, such as "point-stress" or "average-stress" are also used (Kim and Soni,1984). They seem to give good results on specimen under tension, but they relie on weak physical basis. The propagation of an existing delaminated area has been analyzed numerically and experimentally through Fracture Mechanics (Benzeggagh,1980), (Wang,1983). However propagation modelling through Fracture Mechanics uses computations of Energy Release Rates whose values are different from the experimental values under the assumption of perfect bonding between layers (Nesa,1990).

The proposed method is based upon the Damage Mechanics approach proposed in (Ladevèze,1990). The modelling at a structural analysis scale of gradual degradation phenomena in a material - such as T300-914 - due to the occurrence and the development of microcracks, allows an accurate prediction of both initiation and propagation of delamination in a single model. The laminate is modelized as a stacking of single layers and of interfaces connecting the layers. The interface is a two dimensional entity which ensures displacement and traction transfers from one layer to another. Its influence is located near edges or defects where a three dimensional stress state may occur and lead to delamination. Its behavior depends on the angle between the fibre directions of adjacent layers.
Following a "shell" computation, delamination analysis is done in the vicinity of the edge with appropriate boundary conditions. Full analysis of delamination around a circular hole has been developed (Allix,1987). In this work all the non linear phenomena of degradation and plasticity inside the layers and the interfaces between the layers are introduced. This approach allows to take into account possible important gradients parallel to the edge but it needs to solve a three dimensional time-dependant problem which can be expensive with a great number of layers or in the case of parametered studies.

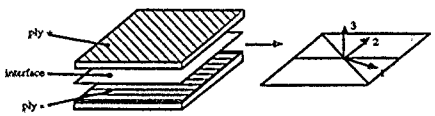
The described approach can be qualified of "simplified" in comparison with the previous one. A Finite Element program has been realized to predict initiation and propagation of delamination at a low numerical cost. The main assumptions of this method are :

- delamination is the main mode of degradation , only the interfaces between the layers can damage
- the edge is weakly curved allowing to consider it is locally straight , the initial three dimensional problem becomes a two dimensional problem set into a band perpendicular to the edge.

First results are presented. The analysis of delamination in the neighborhood of the free edge of a specimen under tension has been treated. This simple example allows comparisons with experimental results (KIM and SONI,1984).

II.LAMINATE MODELLING

The laminate is modelized at the meso-scale, i-e at the scale of the elementary components which are the single layer and the interface connecting adjacent layers.



The single layer - The single layer is homogeneous in the thickness. Its behavior is elastic with no damage.

The interface - The interface is a surface entity (of zero thickness) which ensures stress and displacement transfers from one ply to another (Allix,1987). In elasticity the interface is schematized by means of a model which has been used in order to modelize the fiber-matrix interface (Léné and Leguillon,1981). It depends on the relative orientation of the upper and lower plies.

The displacement discontinuities are denoted by :

$$[U] = U^+ - U^- = [U_1]\,\vec{N}_1 + [U_2]\,\vec{N}_2 + [U_3]\,\vec{N}_3$$

The $(\vec{N}_1, \vec{N}_2)$ axes are associated with the bisector of the fiber directions. The damaged energy of the interface is ($<>_+$ denotes the positive part) :

$$E_D = \frac{1}{2}\left[\frac{<-\sigma_{33}>_+^2}{k^0} + \frac{<\sigma_{33}>_+^2}{k^0(1-d)} + \frac{\sigma_{31}^2}{k_1^0(1-d_1)} + \frac{\sigma_{32}^2}{k_2^0(1-d_2)}\right]$$

k^0, k_1^0, k_2^0 are initial elastic characteristics. If $k^0, k_1^0, k_2^0 \to +\infty$ thus the interface is perfectly bonded, i-e the modelling which is generally used. The conjugate variables associated with the dissipation are :

$$Y_d = \frac{1}{2}\frac{<\sigma_{33}>_+^2}{k^0(1-d)^2} \quad ; \quad Y_{d1} = \frac{1}{2}\frac{\sigma_{31}^2}{k_1^0(1-d_1)^2} \quad ; \quad Y_{d2} = \frac{1}{2}\frac{\sigma_{32}^2}{k_2^0(1-d_2)^2}$$

Mode I — Mode II — Mode III

A simple modeling is to consider that the damage evolution is governed by :

$$\underline{Y} = \sup_{\tau \leq t} (Y_d + \gamma_1 Y_{d1} + \gamma_2 Y_{d2})$$

where γ_1, γ_2 are coupling constants In term of delamination modes, the first term is associated with the first opening mode, the other two with the second and third modes The damage evolution law is defined through a material function w, such that :

$d = w(\underline{Y})$	if $d < 1$	;	$d = 1$ otherwise
$d_1 = \gamma_1 w(\underline{Y})$	if $d_1 < 1$ and $d < 1$	;	$d_1 = 1$ otherwise
$d_2 = \gamma_2 w(\underline{Y})$	if $d_2 < 1$ and $d < 1$	;	$d_2 = 1$ otherwise

A first choice for the material function w is $w(\underline{Y}) = \frac{Y}{Y_c}$ where Y_c is a critical energy.

A first delamination analysis - Link between Damage Mechanics prediction and Fracture Mechanics.

Under the assumption of a damageable interface connecting two elastic layers under a pure mode I loading, an analysis of a Double Cantilever Beam (DCB) specimen has been realized. Because of interface deterioration there exists a maximum value of the applied load P denoted by P_c. This limit is reached for $d=1$ at the tip of the delaminated area. This means that there is delamination propagation This allows us to calculate the critical energy release rate G_{I_c}, that is $G(P_c)$. For a long enough crack one obtains . $G_{I_c} \# 2Y_c$

A similar result has been obtained for the second mode. Within this frame work, Fracture Mechanics appears as a simplified tool to study delamination, under the assumption of elastic layers for an established delamination.

III.NUMERICAL SIMULATION OF DELAMINATION

Geometrical assumption - This study deals with the problem of delamination in the vicinity of a weakly curved edge. Thus the variations of the displacement or stress solution along the direction parallel to the edge are less than those in the perpendicular directions. Therefore the three dimensional problem is reduced into a two dimensional problem in a strip perpendicular to the edge.

Numerical solving - The problem to solve is non linear and time-dependant. It is due to the damage modelling of the interfaces between the plies. The contact between two delaminated plies is described as an unilateral contact without friction. Critical points may occur and then a Riks algorithm is used to detect critical points and to go further with a good convergence.

The numerical treatment for fixed delaminated areas - the problem being purely elastic - is based upon the conjugate gradient method. The problem to solve is :

$$KX = F \quad \text{with} \quad K = K_{lay} + K_{int}$$

K_{lay} and K_{int} stiffness matrix of the layer and of the interface. At each step it is solved :

$$K_{lay}\, V = R \quad \text{with} \quad K_{lay} = \sum_{i=1}^{l} K_{lay} \quad i \text{ denotes the layer}$$

Thus the computation reduces to parallel and separated computations on the layers.

A numerical example - Free edge delamination of a specimen under tension -

The analysis of delamination near the free edge of a specimen T300-5208 (30,-30,90)s submitted to tension has been chosen to illustrate the proposed method. For that example, it is experimentally known that delamination occurs prior to transverse matrix cracking or fiber breaking (KIM and SONI, 1984). Delamination first occurs and propagates on the (90,90) interface, an important damage also appears on the (-30,90) interface. These results confirm the experimental observations.

V.CONCLUSION

A simplified method to analyze delamination near a weakly curved edge has been presented .This engineering tool allows to simulate both initiation and propagation in post-processor of an elastic shell computation through Damage Mechanics. First results are encouraging but it is necessary to make others comparisons with experimental results to valid the model and the numerical program.

VI.BIBLIOGRAPHY

• ALLEN D.H., HARRIS C.E., GROVES S.E., 1987, "A Thermomechanical constitutive theory for elastic composites with distributed damage", International Journal of Solids and Structures, 23-9, pp.1301-1338

• ALLIX O. 1987, "Délaminage par la mécanique de l'endommagement", Calcul des structures et intelligence artificielle, vol 1, Pluralis.

• BENZEGGAGH M.L , 1980, "Application de la Mécanique de la Rupture aux matériaux composites", Thèse de troisième cycle, U.T.Compiègne.

• KIM R.Y., SONI S.R , 1984, "Experimental and analytical studies on the onset of delamination in laminated composites", Journal of Composite Materials, Vol 18.

• LADEVEZE P., LE DANTEC E., 1991,"Damage modelling of the elementary ply for laminated composites", International Journal of Composites Research and Technology (to appear).

• LADEVEZE P., ALLIX O., DAUDEVILLE L., 1990, "Meso-modeling of damage for laminate composites . Application to delamination", IUTAM Inelastic Deformation of Composite Materials, Troy.

• NESA D., 1990, "On cracking in a unidirectional glass-epoxy composite ; Toughness and Damage Mechanisms", International Journal of Fatigue and Fracture of Engineering Materials and Structures (to appear).

• RIKS E., 1981, "Some computational aspects of the stability analysis of non linear structures", Fenomech 81, 2nd International Conference on Finite Elements in Non-linear Mechanics, Stuttgart.

• TALJERA R., 1985, "Transverse cracking and stiffness reduction in composite laminates", Journal of Composite Materials, Vol 19.

• WANG S.S., 1983, "Fracture Mechanics for delamination problems in composite laminates", Journal of Composite Materials, Vol 17.

ON THE PREDICTION BY HOMOGENIZATION OF THE STRENGTH OF DUCTILE COMPOSITE MATERIALS

J. MICHEL and P. SUQUET
Laboratoire de Mécanique et d'Acoustique. C.N.R.S.
31 Chemin Joseph Aiguier
13402 Marseille. Cedex 09. FRANCE.

Abstract: This paper presents a numerical method to predict the macroscopic strength of a composite in terms of the strength of the individual constituents and of their arrangement.

1. THE YIELD DESIGN PROBLEM

In this paper we discuss the strength of composite materials in terms of the strength of individual phases and of their geometrical arrangement. A composite structure, which occupies a domain Ω in $\mathbb{R}^3$, is made of a material which is assumed to exhibit a periodic micro-structure generated by repetition of a unit cell ϵY, where ϵ is a small parameter measuring the size of the heterogeneities and Y is the unit cell (see [1] or [2] for a detailed exposure of these classical notations in homogenization theory). This composite structure is submitted to body forces λf_0 proportional to a loading parameter λ and we search for the admissible values of λ. At the level of the unit cell the strength domain of each constitutive phase, in which the stress tensor is constrained to remain, is known: to simplify we assume that it can be defined by the Von Mises criterion

$$\sigma_{eq}(y) \leq k(y) \text{ where } \sigma_{eq} = \left(\frac{3}{2}\sigma^D_{ij}\sigma^D_{ij}\right)^{1/2},$$

where σ^D is the deviatoric part of σ, $k(y)$ takes constant values, different from one phase to the other. The classical theory of yield design [3] leads to the following dual definitions of the limit load λ^ϵ:

$$\lambda^\epsilon = \sup\left\{\lambda \,\middle|\, \text{there exists } \sigma \text{ such that: } \operatorname{div}(\sigma) + \lambda f_0 = 0 \text{ in } \Omega,\ \sigma_{eq}(x) \leq k\left(\frac{x}{\epsilon}\right) \text{ a.e. } x \text{ in } \Omega\right\} \quad (1)$$

$$= \inf_{\dot u}\left\{\int_\Omega \pi^\epsilon(e(\dot u))\,dx \,;\ \operatorname{div}(\dot u)=0 \text{ in } \Omega,\ \dot u = 0 \text{ on } \Gamma_0,\ L(\dot u) = 1\right\} \quad (2)$$

$$\pi^\epsilon(x,e) = k\left(\frac{x}{\epsilon}\right)e_{eq},\quad e_{eq} = \left(\frac{2}{3}e_{ij}e_{ij}\right)^{1/2},\quad e_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

$$L(u) = \int_\Omega f_0 u\,dx.$$

2. THE MACROSCOPIC STRENGTH DOMAIN

Characterizations analogous to (1) or (2) of $\lambda^{hom} = \lim_{\epsilon \to 0} \lambda^\epsilon$ can be given with the help of the theory of Γ-convergence [4]:

$$\lambda^{hom} = \sup\left\{\lambda \,\middle|\, \text{there exists } \sigma \text{ such that: } \operatorname{div}(\sigma) + \lambda f_0 = 0 \text{ in } \Omega,\ \sigma(x) \in P^{hom} \text{ a.e. } x \text{ in } \Omega\right\} \quad (3)$$

$$= \inf_{u}\left\{\int_\Omega \pi^{hom}(e(u))\,dx;\ u = 0 \text{ on } \Gamma_0,\ L(u) = 1\right\},$$

$$P^{hom} = \left\{\Sigma \in \mathbb{R}^6_s \text{ such that there exists } \sigma:\ \operatorname{div}(\sigma) = 0 \text{ in } Y,\ \sigma.n \text{ is anti-periodic on } \partial Y,\ \langle\sigma\rangle = \Sigma,\ \sigma_{eq}(y) \leq k(y) \text{ a.e. in } Y\right\}, \quad (4)$$

$$\pi^{hom}(\dot E) = \inf_{\substack{u \text{ periodic} \\ \operatorname{Tr}(\dot E + e(u)) = 0}} \langle \pi(y, \dot E + e(u))\rangle,\quad \langle . \rangle = \frac{1}{|Y|}\int_Y . \,dy. \quad (5)$$

π^{hom} is the support function of P^{hom}: $\pi^{hom}(\dot E) = \sup_{\Sigma \in P^{hom}} (\Sigma : \dot E)$.

Expressed in physical terms (3) states that the macroscopic (or homogenized) strength of the composite material is defined by the strength domain P^{hom}, which can be computed, at least theoretically, by (4) or (5). It should be noted that in absence of any further information on the precise constitutive law of the phases, P^{hom} is only an upper bound of the strength capabilities of the composite. However if the individual phases are elastic perfectly plastic, it can be shown that the overall behavior of the composite is elastic plastic (with hardening) and that every stress state Σ in the domain P^{hom} can effectively be reached. For this reason the use of this homogenization theory is permissible for ductile constituents (such as metal matrix composites) but should be considered with care for brittle constituents.

The numerical determination of P^{hom} can be performed directly on the variational definition (5) of its support function, but the functional to minimize has only a linear growth and is not differentiable everywhere. To overcome the difficulty we use another method based on the solution of an elastic perfectly plastic problem [5], for which several FEM codes are available. Fix an arbitrary compliance tensor s and solve the evolution problem:

$$\left.\begin{aligned} &\operatorname{div}(\dot\sigma(y,t)) = 0,\ \langle\dot\sigma(y,t)\rangle = \dot\Sigma(t),\ \sigma.n \text{ anti-periodic on } \partial Y \\ &e(\dot u^*(y,t)) + \dot E(t) = s\dot\sigma(y,t) + \dot e^P(y,t),\ \dot u^* \text{ periodic} \\ &\sigma_{eq}(y,t) \leq k(y),\quad \dot e^P(y,t) = \lambda(y,t)\sigma^D(y,t),\ \lambda \geq 0. \end{aligned}\right\} \quad (6)$$

(6) can be solved along at least two paths. First, it can be solved along a radial path in the space of macroscopic strains: $\dot E(t)$ is constant and prescribed, and (6) is solved for $\dot u^*$, $\dot\sigma$ (from which $\dot\Sigma$ is deduced). Second, (6) can be solved along a radial path in the space of macroscopic stresses: $\dot\Sigma(t)/\Sigma_{eq}(t)$ and $\dot\Sigma(t):\dot E(t)/\Sigma_{eq}(t)$ are constant and prescribed and (6) yields $\dot u^*$, $\dot\sigma$, $\dot E$ and Σ_{eq}. In both cases the solution converges when $t \to +\infty$ and the limits $\dot\Sigma(\infty)$, $\dot E(\infty)$ satisfy:

$$\dot\Sigma(\infty) \in P^{hom},\ (\dot E(\infty) : \Sigma^* - \dot\Sigma(\infty)) \leq 0 \text{ for every } \Sigma^* \text{ in } P^{hom}. \quad (7)$$

Each run of the algorithm determines a point $\dot\Sigma(\infty)$ on the boundary of P^{hom} together with the outer normal vector $\dot E(\infty)$ to P^{hom} at this point.

3. UNIDIRECTIONAL COMPOSITES

Unidirectional composites exhibit geometrical and material properties which are invariant under translation along a specific direction. In the linear setting, this invariance can be used to reduce the 3-d computations to 2-d computations (plane strain and anti-plane problems). In the nonlinear setting considered here the superposition principle is no more valid and the problem remains 3-dimensional in its full generality. However for macroscopic stresses in the form (where e_3 is the invariant direction)

$$\Sigma = \Sigma_{11} e_1 \otimes e_1 + \Sigma_{22} e_2 \otimes e_2 + \Sigma_{33} e_3 \otimes e_3 + \Sigma_{12} (e_1 \otimes e_2 + e_2 \otimes e_1), \quad (9)$$

problem (6) has a solution $\dot{E}$ of the same form (9) and $\dot{u}^*$ is 2-dimensional ($\dot{u}_3^* = 0$, $\dot{u}_1^*$ and $\dot{u}_2^*$ depend on y_1 and y_2 only). (6) is well posed in the context of generalized plane strains. A state of strain is said to be a generalized plane strain if (cf [6]):

$$\dot{u}_\alpha = \dot{u}_\alpha^0(y_1, y_2), \ \alpha = 1,2, \ \dot{u}_3 = \dot{E}_{33} y_3 \ , \text{ where } \dot{E}_{33} \text{ is a constant.}$$

This non classical framework of generalized plane strain gives sufficient flexibility to consider macroscopic stress states with independent components in the longitudinal direction and in the transverse direction, which is not possible with usual plane strain computations.

4. EXAMPLES

4.1 Unidirectional metal matrix composites: We consider a metal matrix composite with unidirectional fibers. The matrix is assumed to be plastic with yield stress σ_0 while the fibers are assumed to be elastic. The fibers are arranged along a triangular array (the unit cell is hexagonal). The yield stress in transverse simple tension has been computed by the above method. Figure 1 shows this yield stress versus the angle between the tensile direction and one axis of the lattice. The reinforcement effect in the transverse direction due to the fibers is rather small and depends only weakly on the volume fraction of the fibers vf. This is due to the presence of planes of maximal shear into the matrix as long as vf does not exceed 0.68.

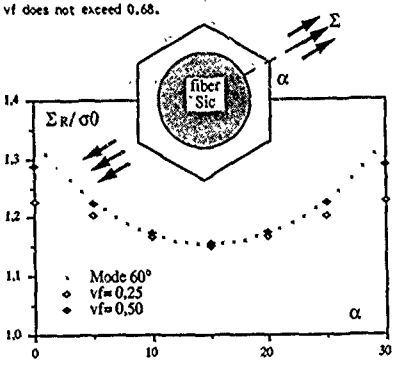
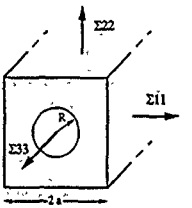


Figure 1: Yield stress of a unidirectional metal matrix composite in transverse tension versus orientation for vf=0,25, vf=0,50

4.2 Cylindrical cavities: We consider a matrix material, with yield stress σ_0, weakened by a periodic array of cylindrical cavities (radius R) arranged along a square lattice (spacing 2a). We investigate the yield locus of this damaged material under macroscopic stresses $\Sigma_{11} = \Sigma_{22} \neq 0$, $\Sigma_{33} \neq 0$, other $\Sigma_{ij} = 0$. Figure 2 shows these yield locii for different values of the pattern ratio R/a. The prediction of the Gurson's model [7] has been reported in dotted line.



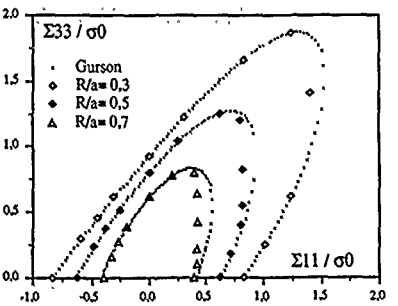


Figure 2: (a) Unit cell, (b) Yield locus of a perforated medium under axisymmetric loading.

REFERENCES

[1] SANCHEZ-PALENCIA E.: Nonhomogeneous media and vibration theory. Springer Verlag. Berlin. 1980.

[2] BENSOUSSAN A., LIONS J.L., PAPANICOLAOU G.: Asymptotic Analysis for Periodic Structures. North Holland. Amsterdam. 1978.

[3] SALENCON J.: Calcul à la rupture et analyse limite. Presses ENPC. Paris. 1983.

[4] BOUCHITTE G., SUQUET P.: "Homogenization, Plasticity and Yield Design". In Composite Media and Homogenization theory. Ed. G. Dal Maso, G. Dell'Antonio. Birkhauser. Boston. 1991. p 107-133.

[5] SUQUET P: "Local and global aspects in the mathematical theory of Plasticity". In Plasticity Today. Ed. A. Sawczuk and G. Bianchi. Elsevier. Amsterdam. 1984. p 279-310.

[6] LICHT C., SUQUET P.:"Augmented Lagrangian method to a problem of incompressible viscoplasticity arising in homogenization". In Numerical methods for Nonlinear Problems. Ed. Taylor et al. Pineridge Press. Swansea. 1986. p 106-114.

[7] GURSON A.L.:"Continuum theory of ductile rupture by void nucleation and growth: I. Yield criteria and flow rules for porous ductile media". J. Eng. Mat. Tech., 1977, 99, p 1-15.

Some mixed finite element methods for composite plaques .

Jean-Luc Akian
O.N.E.R.A.
29 Av. de la Division Leclerc
BP 72 92322 CHATILLON Cedex
FRANCE

Abstract .

The mixed FEM proposed in order to compute accurately the 3D stress field in a composite plate are based on modifications of the principle of the complementary energy. The main difficulty of mixed methods is to construct admissible fields such that the final linear system is invertible and the method converges .

We first use the 3D complementary energy principle and we suppose that the in-plane stresses are linear functions in each layer(this is the cruder physical hypothesis we can do).As the admissible stress fields must satisfy the equilibrium equations in the composite,we get a 2D formulation with 16 generalized stress variables per layer, defined on the middle plane of the composite (ω).

Let Θ_h be a family of triangulations of ω. In order to satisfy the equilibrium equations,we have to construct symmetrical tensors $T = (T_{\alpha\beta})_{\alpha,\beta=1,2}$ and fields $q = (q_\alpha)_{\alpha=1,2}$ such that :

$$q \in (L^2(\omega))^2 \ , \ \mathrm{div}\, q \in L^2(\omega) \qquad (1)$$

$$T \in (L^2(\omega))^4_s \ , \ \mathrm{div} T \in (L^2(\omega))^2 \ , \ \mathrm{div}(\mathrm{div} T) \in L^2(\omega) \ . \qquad (2)$$

In order to satisfy (1),we can use the Raviart-Thomas space ([2]) :

$$D_k(K) = (P_{k-1}(K))^2 + \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} P_{k-1}(K) \ , \qquad (3)$$

(K is a triangle, k is an integer ≥ 1 , $P_{k-1}(K)$ is the space of polynomials of degree $\leq k-1$ defined on K).

In order to satisfy (2), we can use the following lemma:

Lemma

Let T belong to $(P_2(K))^4_s$, then T is uniquely determined by :

a) the moments of order 0 of $T_{\alpha\beta}$ on K ($\alpha,\beta=1,2$) ,

b) the moments of order 0,1,2 of $T_{\alpha\beta} n_\alpha n_\beta$ on the edges of K ,

c) the moments of order 0,1 of $T_{\alpha\beta} n_\alpha t_\beta$ on the edges of K ,

($n = (n_\alpha)_{\alpha=1,2}$, $t = (t_\alpha)_{\alpha=1,2}$ are the unit normal and tangent vectors along the edges of K) .

If we choose the spaces :

$$E_{\alpha\beta} = \left\{ T \in (L^2(\omega))^4_s \ / \ T_{|K} \in (P_2(K))^4_s \ , \ K \in \Theta_h \right\} \qquad (4)$$

$$E_{3\alpha} = \left\{ q \in (L^2(\omega))^2 \ / \ q_{|K} \in (D_1(K))^2 \ , \ K \in \Theta_h \right\} \qquad (5)$$

$$E_{33} = \left\{ u \in L^2(\omega) \ / \ u_{|K} \in P_0(K) \ , \ K \in \Theta_h \right\} \qquad (6)$$

for the variables related to $\sigma_{\alpha\beta}$, $\sigma_{3\alpha}$, σ_{33} , and if we use Lagrange multipliers to satisfy the continuity of the d o f. 'normal traces' characterizing the spaces $(P_2(K))^4_s$ and $D_1(K)$ we can prove that the final linear system is invertible .

The admissible spaces can be enlarged. Let

$$(\phi_1)_{\alpha\beta} = x_1^{2-(\alpha+\beta)} \, x_2^{(\alpha+\beta)-2}$$

$$(\phi_2)_{\alpha\beta} = x_1^{4-(\alpha+\beta)} \, x_2^{(\alpha+\beta)-1} \quad (\alpha,\beta=1,2) \ ,$$

if we replace

$(P_2(K))^4_s$	by	$(P_2(K))^4_s + R\phi_1 + R\phi_2$	in (4)
$D_1(K)$	by	$D_2(K)$	in (5)
$P_0(K)$	by	$P_1(K)$	in (6)

(method number 2)

or

$(P_2(K))^4_s$	by	$ADG(K) + R\phi_1 + R\phi_2$	in (4)
$D_1(K)$	by	$D_2(K)$	in (5)
$P_0(K)$	by	$P_1(K)$	in (6)

(method number 3)

(here ADG(K) is the Arnold-Douglas-Gupta space[1]), after slightly modifying the lemma and the Lagrange multipliers, the same conclusion holds.

In method number 3,we are able to construct 3D stress fields verifying exactly the equilibrium equations in the composite .

If we simplify the formulation in method number 2 by considering only plane elasticity problems, or in method number 3 by considering only bending of plates problems, we can prove convergence results.

As far as the implementation of the method is concerned, all the d.o.f. related to the stresses can be eliminated locally. As the d.o.f. on the edges of the triangles are related to 2 elements we can divide the mesh in a few clusters of elements and use an element by element iterative method, in order to solve the final linear system. This technique is very efficient on parallel/vector supercomputers .

Some tests have been made (plane elasticity problems,bending of plates problems),and the results are all excellent, on both mechanical and computational efficiency points of view .

[1] Arnold D.,Douglas J.,Gupta C., *"A family of higher mixed finite element methods for plane elasticity"*, Num. Math.-45-1984 .

[2] Thomas J.M., Thesis, Universite de Paris VI-1977 .

Parallel implementation of a domain decomposition method for composite three dimensional structural analysis problems .

François-Xavier Roux ,
O.N.E.R.A.
29 Av de la Division Leclerc
BP 72 92322 CHATILLON Cedex
FRANCE
E.mail roux@onera.fr

Abstract .

Domain decomposition methods have very interesting features to solve three dimensional composite structural analysis problems. Indeed, the composite feature of the material leads to a natural mesh substructuring, with subdomains that have similar stiffness matrices, that entails optimal efficiency with domain decomposition solvers on parallel computers .

In this paper, we present some tests performed with a domain decomposition method via Lagrange multipliers, developped in [1], on a distributed memory parallel computer, the Intel iPSC2 .

The method is based upon a hybrid variational formulation of the linear elasticity equations that consists in enforcing the continuity constraint along the interface between the subdomains by introducing the Lagrange multiplier of this constraint .
The Lagrange multiplier appears to be equal to the interaction force between the subdomains .

The solution process consists in solving the condensed problem related to the Lagrange multiplier on the interface by the conjugate gradient algorithm. So, the condensed interface operator does not need to be actually assembled, and the computation of the matrix-vector product involves the solution of local independant problems in each subdomain. The use of a direct skyline solver in the subdomains allows a good vector efficiency .

As the Lagrange multiplier just enforces the continuity constraint, the Ladyzenskaia-Babuska-Brezzi condition does not need to be satisfied, as far as the discrete displacements fields are continuous at the interface .

We analyze the performance obtained with different splittings for a three dimensional composite cantilever beam, on both numerical and parallel processing points of view. These tests prove that this method performs much better than the classical substructures method. At last we compare the results obtained with various implementation strategies in order to define the optimal one .

References .

[1]Ch. Farhat and F.-X. Roux , "An Unconventional Domain Decomposition Method for an Efficient Parallel Solution of Large-Scale Finite Element Systems ," to be published in *SIAM Int. J. Sci. Statist. Comput.* .

Boundary element techniques for composite materials.

by

F. LENE and P. PAUMELLE

Groupe Mécanique, Modelisation et Calcul
Laboratoire d'Acoustique et Mécanique
Université Pierre et Marie Curie
Tour 66
4, place Jussieu
Paris Cedex 05

Abstract

The homogenization techniques are particulary powerful for studying and modelizing the composites materials behaviour. They lead to the resolution of problems depending on a basic cell which characterizes the medium studied. These problems are usually solved by use of a finite element method. In this study we show how to solve the cell problems by using the boundary integral method. Contrary to classical finite element methods formulated in displacements, the unknowns are here the fields of displacements and stress vector on the boundary of the basic cell and at the interfaces between its components. Thus, the boundary integral method appears as a well suited tool to analyse the damage phenomenons which may appear at the interface.

Introduction

The main difficulty with studying composite materials is their high level of heterogeneity compared with the dimensions of the structures, making any numerical computation (for example, by finite element method) prohibitive if not impossible. The technique commonly used to get around this obstacle consists of substituting for this highly heterogeneous material a homogeneous one having "average" mechanical properties that depend on the mechanical properties and geometry of the various constituents. This is the homogenization process (Duvaut [4], Sanchez Palencia [11]). Furthermore, by a localization procedure, the method allows an easy computation of the microscopic field of stresses and, in particular, of stress forces at the boundaries between heterogeneities and matrix. One can then show the overstresses phenomenon may initiate decohesions at the microscopic level.

In the first part of this paper we show the F.E.M. results induced by these techniques. This application refers to the example of the woven composites (Paumelle [9]). In order to obtain a good approximation of the local fields, the cost of the computations may be expensive in this case.

The second part is devoted to the implementation of the boundary integral method (Mikhlin [8], Brebbia [2], Dautray [3], Hartmann [6]) for homogenization. We show that it allows to obtain a good approximation of the stress vector. Thus, it may be a useful tool to develop models which take in account a damage localized at the interfaces between the components of the basic cell.

1. Homogenization by F.E.M. : application to woven composites

A lot of computations about composites materials (unidirectional composites, syntactic foams, ..) have been made in our laboratory. They have been realized by use of F.E.M. and the code MODULEF. When it is a matter of considering composites whose basic cell needs to be tridimensional (as woven composites), the induced computations may be expensive. Moreover our aim being to predict the sensitivity to damage near the interface of the components, the mesh has to be sufficiently refined in order to obtain a very good approximation of the stress vector.

As the discretization of these problems is made by use of linear pentahedral and hexahedral Lagrange finite elements, the stresses we obtain are constant over each of the elements. So, it is difficult to obtain a satisfying approximation of the interface stress vector. This one was nevertheless well computed by use of the numerical technique developed by G Duvaut and F. Pistre [5]. To this way we have used the adapted post processor developed in our laboratory.

Some results presented in figure 1 show the norm of the tangential traction introduced by an imposed macroscopic stress $\langle \sigma^{12} \rangle = 100$ Mpa. An analyse of the results we have obtained, allows a prediction of the damage sensivity by sliding or decohesion at the interfaces.

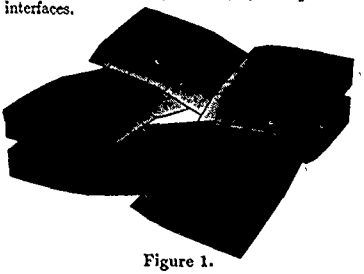


Figure 1.

- A damage model by sliding.

In order to modelize a sliding phenomenon between the yarn and the matrix, which is often experimentally observed, we have introduced a behaviour's law of the interface allowing a tangential elastic sliding without normal unstiking (Léné, [7]) It writes :

$$\tau_T = k[[v_T]]; \quad [[v_N]] = 0$$

where $[[(v_i)_T]]$ (respec $[[v_N]]$) is the tangential jump of the displacement (normal jump) and τ_T the tangential stress vector at the interface.

This phenomenon has been taken in account inside a software we have developed.

We present on figure 2. the values of the homogeneous equivalent modulus versus the parameter of sliding k It appears that there exist limit values noted k_m and k_M of this scalar which determine three regions. The first ($k \geq k_m$) corresponds to a behaviour without sliding; the third ($k \leq k_M$) to a total decohesion between the yarn and the resin; the middle region ($k_m \leq k \leq k_M$) to a partially debonded yarn.

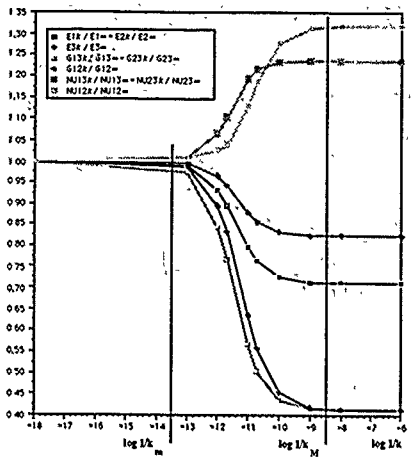


Figure 2.

In order to develop models more sophisticated than the previous one in the framework of the F.E.M., we should be led to consider complex problems with expensive resolutions. Our aim is to develop realistic damage models at the interface, constructed above the stress vector. Thus, it appears necessary inside the framework of the homogenization to develop a more adequat method.

2. Integral equations for homogenization

To describe the homogenized behaviour of composite materials, we have used the integral formulation of the cell problems (Paumelle [10]). The numerical method associated with the integral formulation is called *Boundary Elements Method*. The unknown functions of the formulation are the displacement u and the stress vector $\tau(n)$ on the boundaries. These one are the external side of the basic cell and the interface between the components.

In order to numerically implement this technique, we have constructed three bidimensional elements with a constant, linear and quadratic interpolation and one three dimensionnal constant element. One can note that in the case of homogenization we have introduced some specific boundary conditions and manage the presence of some domains which need a special numerical treatment.

The results shown here are related to unidirectional composite materials and are carried out on a SUN 4 workstation. As one can see they are quite satisfying The approximations of the homogenized coefficients, the microfields and the stress vector on the interfaces are as good as the same results obtained by finite element method. One test is interesting to discern the difference obtained between constant, linear and quadratic interpolations It consists in studying the convergence speed of the homogenized coefficients versus the number of the nodes of the mesh (Fig 3) It shows that the results with constant elements are as much powerful than the others. One can notice that over 200 nodes the results tend to the same values which are the finite element values.

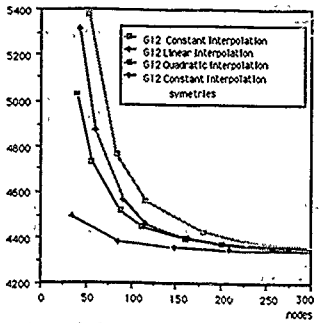


Figure 3.

In conclusion, this technique appears as a good way to characterize the behaviour of composites and to obtain the microme chanical properties. The analyse of results obtained by F.E.M shows a nearly perfect agreement (less than 1.3 %) with those obtained by the boundary integral method. The agreement of the results is nearly perfect .

This study has emphazed the high level of quality of constant element. Moreover, we have seen that the number of nodes doesn't need to be high. Thus, the number of degrees of freedom stay weak. One can also note that it is very easy to add some nodes in a part of the mesh (interface for instance) without altering the whole mesh. And the directly computation of the stress vector on the interfaces presents a real advantage for a further analyse of the microphenomenon of damage.

References

[1] D. K. Banerjee and D. K. Butterfield, *Boundary Element Methods in Engineering Science*, Mac Graw-hill, (1981).

[2] C A. Brebbia, J.C.F. Telles and L.C. Wrobel, *Boundary Element Techniques*, Springer-Verlag, (1984).

[3] R. Dautray and J.L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson, Collection CEA, Tome 2, (1985).

[4] G. Duvaut, *Analyse Fonctionnelle et Mécanique des Milieux continus, Application à l'étude des matériaux composites élastiques à structure pénodique, Homogénéisation.*, Theorical and Applied Mechanics, W.T. Koiter éd., North Holland Publishing Company, (1976).

[5] G. Duvaut and F. Pistre, *Calcul des vecteurs contraintes en approximations P1 et P2.*, C.R.A.S., T. 295, Serie II, pp 827, (november1982).

[6] F. Hartmann, *Introduction to Boundary Elements*, Springer-Verlag, (1989).

[7] F. Léné, *Contribution à l'étude des matériaux composites et de leur endommagement*, Thèse d'état, U.P.M.C., (1984).

[8] S.G. Mikhlin, *Multidimensional Singular Integrals and Integral Equations*, Pergamon Press, (1965).

[9] P. Paumelle, A. Hassim and F. Léné, *Les composites à renforts tissés . Calcul et Etude Paramétrée du Comportement Homogénéisé*, La Recherche Aérospatiale, vol. 1, p. 1 (1990).

[10] P. Paumelle, *Homogénéisation de matériaux à structures pénodiques par les méthodes intégrales*, Textos e Notas, CMAF, Lisbonne, (1991).

[11] E. Sanchez - Palancia, *Non homogeneous media and vibration theory*, Lecture notes in physics, number 127, Springer-Verlag, Berlin, (1980).

SOME ATTEMPT TO BRIDGE THE MACRO-AND MICROSCOPIC ASPECTS

Tadahiko KANAI

Department of Electrical Engineering, Science University of Tokyo 162 Japan

Abstract

Finite element Method is very powerful method for analysis of nonlinear problems of solids and structures.

Unfortunately, however, it requires a great amount of manpower and computing time and to make it worse accuracy of the solution can hardly be guaranteed.

To solve such difficult problems, the present author developed a new discrete model entitled "Rigid Bodies -Spring Models" in 1976 by using which the so called limit analysis may be generalized including effects of instability as well as cracking so that failure load analysis of any solid material may be possible.

Since then this discrete method of limit analysis has been duly verified by quite a few numerical examples on the failure load analysis of metal, concrete as well as foundation structures.

12 papers will be presented in two sessions entitled "Recent Developments on the Discrete Method of Limit Analysis of Solids in Japan".in this Congress.

1.INTRODUCTION

Nature is essentially discrete and sometimes we observe its unstable and discontinuous behavior such as called "Catastrophe"or "Chaos".

It seems very difficult to understand these phenomena completely in the framework of so-called continuum physics or continuum mechanics.

They are generally called the nonlinear problems and we are obliged to tackle them in every frontier of science and technology.

It is the author's belief that they are all due to the mathematical penalty created in understanding the discontinuous phenomena from the viewpoint of the continuum physics or mechanics.

There must be two ways to grapple with these problems; one is the traditional approach of the applied mathematics, applied mechanics, computational topology and the other will be the simulation of the modern digital computers.

Judging from the remarkable progress of the computational physics and mechanics, we may go much further beyond we are now in near future.

But the present author is rather pessimistic about their future of this approach unless some moving discontinuities can be successfully treated.

He believes that there are two faces in nature; one can be easily explained by the continuum approach, while the other face can be only understood by the discrete approach.

Theory of light is a typical example.

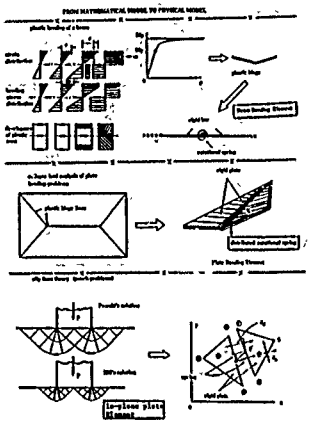
Therefore, new fruitful development in sciences and technology could be anticipated by beautiful computer simulation using elaborate discrete models in future.

2.Recapitulation of the Theory of RBSM

In order to break the present deadlock of finite element nonlinear analysis in solid mechanics, the present author proposed a family of new simplified elements in 1976 based on the experimental evidence of solids under the ultimate state of loading as shown in Fig.1.

These elements consist of rigid bodies and two types of connection springs, one of which resists the dilatational movement, while the other resists the sliding movement.

(See Fig.2)



Establishment of the mathematical basis of these elements has been recently successful to some extent.

Generally granular materials such as soils and rocks are too inhomogeneous, nonuniform, and easy to slip internally under applied loading to consider them as continua.

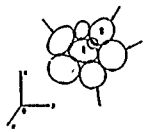
Consequently the method of limit analysis has been introduced many years ago in this area and it has proved to be useful, but its application is still limited due to many reasons.

Takeuchi has developed a new method of limit analysis on the mechanics of granular materials which have low tensile strength by using the RBSM models.

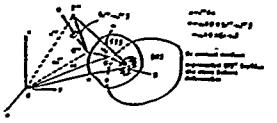
In this method he developed a checking scheme for crack initiation in the interelement spring system and recontact of cracked surfaces.

And he introduced it to "No-Tension Analysis" which was proposed by Zienkiewicz and others to study problems of soil and rock mechanics as shown in Fig.3.

Through a series of numerical studies on typical problems in soil and rock mechanics validity of "Method of Tension Crack Analysis" was duly verified and especially its usefulness was proved by analysis of solid contact problems for which the finite element method is experiencing serious difficulty.



(a) Idealization of a three dimensional body as assembly of 3D RBSM models



3) Displacement vector in the RSSM system

Rigid body displacement field (i th element)

$U_i=u_i+\omega\times(r-r_i)=B_i a_i$

$(r-r_i)$: position vector of an arbitrary point with respect to the centroid

U_i: the displacement vector of an arbitrary point of a given element

a_i: rigid body displacement vector of the element centroid $(u_i, v_i, w_i, \theta_i, \gamma_i, \zeta_i)$

u_i, ω: translational and rotational displacements of the element centroid respectively

$a_i=[u_i, v_i, w_i]$, $\omega=[\theta_i, \gamma_i, \zeta_i]$

B_i: shape function of the rigid body displacement field relative displacement vector on the interelement boundary

where $\delta=Bu$

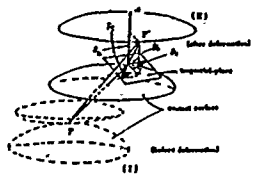
$\delta=U_j-U_i$, $u=(u_i, u_j)$

strain energy to be stored in the boundary spring system

$$V=\frac{1}{2}\iint(\lambda\delta_n^2+k\delta_s^2)dS=\frac{1}{2}\iint\delta^T D\delta\, dS$$

$$=\frac{1}{2}u^T Ku\left(K=\iint B^T DB\, dS\right)$$

where k_n, k_s are spring constants of the normal and tangential spring system respectively



(I)

Rubber deformation

Tangential plane

Contact surface

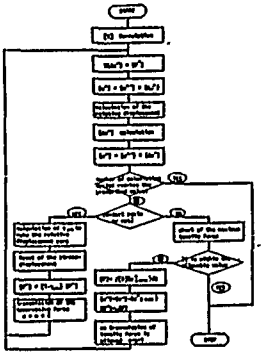
Rubber deformation

(II)

δ_n: relative normal displacements of two rigid elements (I)-(II) (dilatation) on the point P.

δ_s: tangential relative displacement of elements (I)-(II)

(a) Contact boundary surface of elements (I)-(II) (enlarged figure of the domain encircled in (a))



3. Recent Progress of RSSM Developments

For the last five years development of the RSSM elements has been continuously made by colleagues of the present author in various fields of structural mechanics.

In this congress he organized two sessions entitled "Recent Development on the Discrete Method of Limit Analysis of Solids in Japan" where 12 papers will be presented.

Four papers are concerned with ultimate strength of metal there walled members, while five papers will discuss geotechnical problems, the other three will treat the load carrying capacity of the reinforced concrete structures.

4. GENERALIZATION OF RSSM MODELS.

RSSM models are originally proposed to study the macroscopic behavior of matters.

It is clear that they are identical to Cundall models when no-tension springs are assumed on the interelement boundaries.

If the spring connecting two elements (or particles) is replaced by reaction forces which may be derived from their potentials, a variety of new discrete models can be derived.

Let us consider a system of rigid ellipsoidal bodies whose potentials are given by the following equations;

$$\phi(r_{ij})=\sum_i\sum_j a_{ij}^{(n)} r_{ij}^{-2n}$$

where r_{ij} is the distance between centroids of two elements.

If $\phi(r_{ij})$ is truncated up to the second order term, i.e.

$$\phi(r_{ij})=\sum_i\sum_j (a_{ij}^{(1)}-a_{ij}^{(2)} r_{ij}^{-2})$$

the corresponding motion of bodies represents the famous "many body problem" in celestial mechanics.

When $\phi(r_{ij})$ is given by the following equation;

$$\phi(r_{ij})=\sum_i\sum_j (a_{ij}^{(2)} r_{ij}^{-6}+a_{ij}^{(1)} r_{ij}^{-12})$$

these potential obviously represent the well-known van der Waals dispersion forces in molecular dynamics, etc.

Further details on the derivation of governing equations and their time integration by the incremental procedure will be discussed at the session of this congress.

REFERENCES

1. Nagtegaal, J. C., Parks, D. M., and Rice, J. R., "On Numerically Accurate Finite Element Solutions in the Fully Plastic Range", *Comp. Meth. Appl. Mech. Eng.*, 4, 1974, pp. 153-178.

2. Zienkiewicz, O. C., *The Finite Element Method*, McGraw-Hill Book Co. (U. K.), 3rd ed., 1977.

3. Zienkiewicz, O. C., et al., "Stress Analysis of Rock as 'No-tension' Material", *Geotechnique*, 18, 1968, pp. 56-66.

4. Kawai, T., "New Discrete Structural Models and Generalization of the Method of Limit Analysis", International Conference on Finite Elements in Nonlinear Solid and Structural Mechanics, Geilo, Norway, August 29-September 1, 1977, Vol. 3, C04.1-C04.20.

5. Kawai, T., "New Discrete Models and Their Application to Seismic Response Analysis of Structures", *Nuclear Engineering and Design*, Vol. 48, 1978, pp. 207-229.

6. Kawai, T., and Toi, Y., "A New Discrete Analysis on Dynamic Collapse of Structures", *Journal of the Society of Naval Architects of Japan*, Vol. 142, May 1978, pp. 275-281, Vol. 145, May 1979, pp. 112-119.

7. Kawai, T., "A New Approach to Soil Mechanics and Geotechnical Engineering", Keynote address to the Third International Conference in Australia on Finite Element Methods, July 2-6, 1979, Sydney, 1979

8. Kawai, T., Yokouchi, N., and Kanio, Y., "A New Discrete Limit Analysis of Underground Structures", Recent Advances in Lifeline Earthquake Engineering in Japan, ASME Publication PVP-43, Aug. 1980.

9. Kawai, T., "Some Considerations on the Finite Element Method", *International Journal for Numerical Methods in Engineering*, Vol. 16, 1980, pp. 81-120.

FINITE ELEMENT ANALYSIS OF HIGHLY NONLINEAR BEHAVIORS OF STEEL STRUCTURES

Yutaka Toi
Institute of Industrial Science, University of Tokyo,
7-22-1, Roppongi, Minato-ku, Tokyo, 106 Japan

Kohei Yuge
Faculty of Engineering, University of Seikei
3 Cho-me, Kichijoji-Kitamachi, Musasino City,
Tokyo, 180 Japan

Abstract

The highly nonlinear behaviors of steel structures are simulated by the finite element method which employs a bilinear degenerated shell element with the reduced integration technique. The calculated problems are as follows : 1) crashworthiness of axially compressed square tubes with various thickness ratios 2) ultimate strength of tubular column to H-beam connections 3) reinforcement effect on transverse collision collapse of box beams. The obtained results are compared with the experimental results or the empirical formula in order to discuss the quantitative validities of the present analysis.

1. Outlines of the Present Finite Element Code [1]

The present finite element code for the quasi-static large deformation analysis is based on the following algorithm : 1) incremental theory by the updated Lagrangian approach 2) zero normal-stress projection [2] 3) orthogonal hourglass control [3] 4) additional stiffness resisting in-plane rotation [4]. In the present formulation, the updated Kirchhoff stress increments calculated at each loading step are transformed into Jaumann rates of Euler stresses. 'Zero normal stress projection' is the technique to maintain the plane stress condition during this transformation. It should be noted that the experimental true stress-true strain relation must be used as the constitutive equation.

2. Crashworthiness of Axially Compressed Square Tubes [5]

Finite element analysis for crashworthiness of square tubes under axial compression has been conducted. A square tube under axial loading collapses with a progressive buckling, but it will cost too much to simulate this phenomena directly. Therefore, assuming that progressive bucklings occur simultaneously on the tube, the mean crushing stress for a unit periodic area of buckling patterns has been calculated with a minimum computing cost. In Fig. 1, the calculated mean crushing loads expressed in terms of the tensile strength are compared with the experimental values for various thickness to edge length ratios. The bold line in the figure is the empirical formula given by Magee et al. [6]. The calculated mean crushing stresses are about 25 % smaller on the average than the experimental values. This difference is caused by the above assumption as well as the limitation of 'thin walled assumption', because the crush analysis is always accompanied with fairly larger strains than those in the ordinary ultimate strength analysis. However, there is a good agreement between the calculated and experimental results from a qualitative point of view. Fig. 2 and Fig. 3 are the experimental and the numerical crushing deformations respectively, which agree well with each other.

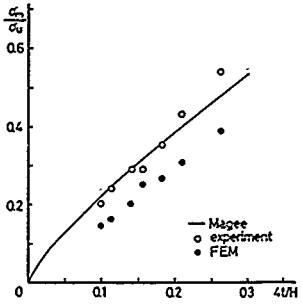


Fig. 1 Mean crushing stresses of axially compressed square tubes

Fig. 2 Axially compressed square tubes

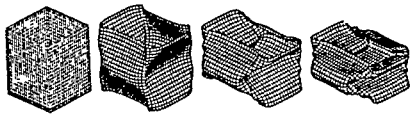


Fig 3 Calculated crushing deformations of axially compressed square tubes

3. Ultimate Strength of Tubular Column to H-beam Connections [7]

Influences of a diameter-thickness ratio of a circular column on the ultimate strength and the collapse mode have been studied. Dimensions and boundary conditions for the calculated connection are shown in Fig. 4 and Table 1. Calculated four specimens have different diameter-thickness ratios for the column. The half region of the connections was subdivided into elements. The calculated load-displacement curves are shown in Fig. 5. The predicted maximum loads by the empirical formula [8] for Case 1 and Case 2 are 34.4 tons and 47.9 tons respectively which agree well with the calculated values. In Fig. 6 the calculated collapse modes are shown. Depending on diameter thickness ratios, the four specimens have collapsed in different modes. In Case 1 and 2 where columns are relatively thin, the circular columns locally buckled. In Case 3, both the flanges of the H-beam and the circular column yielded at the same time. Therefore, it can be said that the strength of the circular column and the H-beams are well balanced in Case 3. In Case 4 where the column is relatively thick, the torsional-flexural buckling with a local buckling of the flanges have occurred, while the column has deformed little. The present method would be useful for the optimal design of these connections.

4. Reinforcement Effect on Transverse Collision of Box Beams [9]

The reinforcement effect on the transverse collision collapse of box beams has been studied. A spot welded(Case 3) or a line welded(Case 4) reinforcement plate is attached on the colliding surface of box beams and their strength has been compared with that of box beams with no reinforcement plates. The loading apparatus and the calculated box beam are shown in Fig. 7 and Fig. 8, respectively. The box beams are assumed to collide with a rigid circular column at a speed of 25 km/h. In this section, the quasi-static algorithm used in the former sections is extended to the dynamic transient analysis based on the explicit time integration scheme. In addition, the influence of a strain rate is taken into account by the elasto-viscoplastic flow theory, and the nodal degrees of freedom are shifted from the element mid-surface to the outer surface at the welded points. The obtained load-displacement curves are compared with experimental

Table 1 Dimensions of calculated connections

Case No.	$D_p/_pt_p$	$_pt_p$ (cm)	D_p (cm)	t_p (cm)	t_s (cm)	h_s (cm)	R^* (cm)
Case 1	53.09	0.55	29.20	0.379	1.52	3.45	14.325
Case 2	41.93	0.70	29.35	0.483	1.93	3.38	14.325
Case 3	20.76	1.45	30.10	1.00	4.00	3.00	14.325
Case 4	10.55	3.0	31.65	2.07	8.28	2.23	14.325

*: $R=(D_p-{}_pt_p)/2$

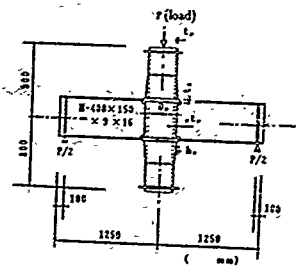


Fig. 4 Tubular column to H-beam connection

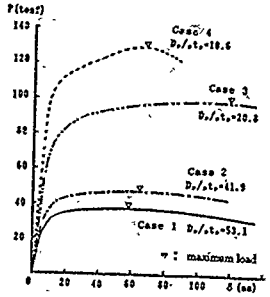
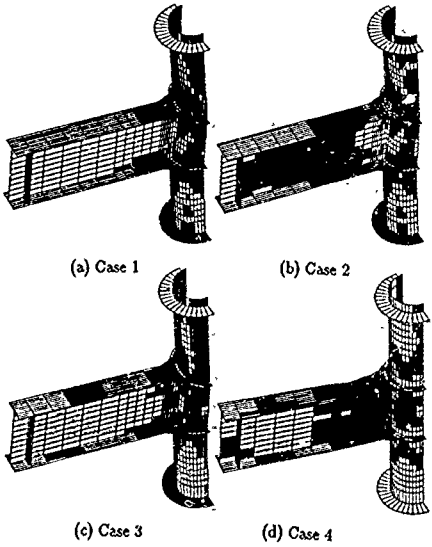


Fig. 5 Load-displacement curves for the connections



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Fig. 6 Collapse modes for the connections

maximum loads in Fig. 9. It is obvious that the present analysis explains the difference of the ultimate strength of four tested specimens.

5. Concluding Remarks

The numerical results for highly nonlinear behaviors of steel structures subjected to large deformations have been briefly described. It is obvious that the finite element code developed in the present study can be a powerful tool for the basic study of crush phenomena as well as the optimum design of actual structural components. Other results are published in [1],[5] and [10].

References

[1] Y. Toi, K. Yuge and T. Kawai : Finite Element Analysis of Axially Compressed Circular Cylindrical Shells, Proc. of the Int. Conf. on Computational Mechanics, (1986), 4-217.

[2] T. J. R. Hughes and W. K. Liu : Nonlinear Finite Element Analysis of Shells (Part 1. Three-dimensional shells), Comp. Meth. in Appl. Mech. and Engng., Vol.26, (1981), 331.

[3] D. P. Flanagan and T. Belytschko : A Uniform Strain Hexahedron and Quadrilateral with Orthogonal Hourglass Control, Int. J. Num. Meth. Engng , Vol. 18, (1981), 167.

[4] W. Kanoknukulchai : A Simple and Efficient Finite Element for General Shell Analysis, Int. J. Num. Meth. in Engng., Vol.14, (1979), 179.

[5] Y. Toi, K. Yuge, T. Nagayama and K. Obata : Numerical and Experimental Studies on the Crashworthiness of Structural Members, Naval Arechitecture and Ocean Engineering, Vol. 26, (1988), 91.

[6] C. L. Magee and P. h. Thornton : Design Considerations in Energy Absorption by Structural Collapse, S.A.E. Paper, No. 780434 (1978)

[7] K. Yuge, Y. Toi and M. Teraoka : Finite Element Analysis of Ultimate Strength of Steel Tubular Column to H-Beam Connections, Proceedings of Symposium on Computational Methods in Structural Engineering and Related Fields Vol. 13(1989), 371. (in Japanese)

[8] T. Kamba, H. Kanatani, M. Tabuchi and T. Wakida : Local Strength of the Centrifugal Cast Steel Tubular Columns to H-Beam Connections, J. of the Soc. of Architecture of Japan, Vol. 370(1986), 81. (in Japanese)

[9] K. Yuge and Y. Toi : Finite Element Analysis of Reinforcement Effect on Transverse Collision Collapse of Box Beams, Proceedings of Symposium on Computational Methods in Structural Engineering and Related Fields Vol. 14(1990), 121. (in Japanese)

[10] Y. TOi, K. Yuge, T. Nagayama and K. Obata : Finite Element Crush Analysis of Structural Components and Experimental Validations, Proc. of the Int. Pacific Conf. on Automotive Eng., (1987).

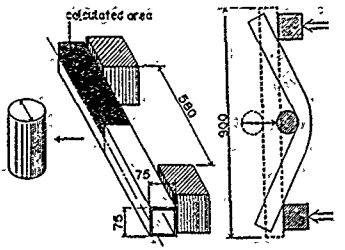


Fig. 7 Transversely colliding box beam

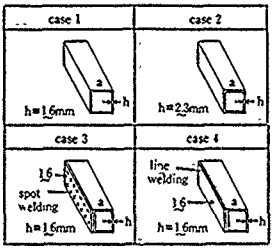


Fig. 8 Calculated box beams

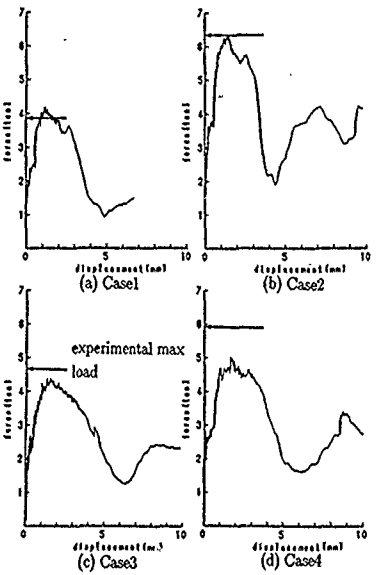


Fig. 9 Load-displacement curves for the box beams

NUMERICAL ANALYSIS OF TECTONIC DEFORMATION BY MAGMA INTRUSION AND PREDICTION OF VOLCANIC ERUPTION

HITOSHI KOIDE
Geological Survey of Japan
Higashi 1-1-3,Tsukuba
Ibaraki 305,Japan

RYOUKICHI HAMAJIMA
Saitama University
255 Shimookubo,Urawa
Saitama 338,Japan

TADAHIKO KAWAI
Science University of Tokyo
1-3,Kagurazaka,Shinjyuku
Tokyo 162,Japan

and MANABU TAKAHASHI
Geological Survey of Japan
Higashi 1-1-3,Tsukuba
Ibaraki 305,Japan

Abstract-Numerical analysis by discrete models was used to reveal the shape and depth of underground magma body from pattern of ground movement of the Izu-Oshima volcano, Japan. Measurement of areal pattern of ground movement of volcanic surface is useful for tracing of underground movement of magma body and for prediction of volcanic eruption.

I INTRODUCTION

Discrete models are suitable for numerical simulation of tectonic deformation of the Earth's crust which is essentially discontinuous with faults, joints and various boundaries of rocks. The Japanese Islands locate near of boundaries between major plates:Eurasia plate, Philippine Sea plate, Pacific plate and North American plate. Active interaction between plates induces frequent earthquakes and volcanic activities, in Japan. On the November 15th, 1986, the first volcanic eruption since 1974 started in the summit crater of the Izu-Oshima volcano, Japan. Earthquake swarms, volcanic tremors, etc. signaled the earlier subsurface magmatic activity, but the subsidence of central surface of the volcano prevented the short-term prediction of the eruption as people believed that upraise of subsurface magma body induces upheaval of ground, only.

In many cases, the inflation of central surface of volcano proved of the most reliable precursor of volcanic eruption. However, Koide and Bhattacharji(1975) pointed out that rock mass could be depressed in the central area by the increase of magma pressure above the magma body which has long-vertical depth with narrow horizontal section.

II NUMERICAL ANALYSIS BY DISCRETE MODELS

In this paper, the authors applied the rigid-body-spring method (RBSM) by Kawai(1980) for a realistic simulation model of tectonic effect of subsurface magma intrusion where differential movements along fractures have vital role for deformation of the crust. In the two dimensional calculation, an elliptical hole is assumed in a rectangular plate(Fig.1). The wall of hole is applied with magma pressure. At the start of calculation, magma pressure is equal to the lithostatic pressure at the top of magma reservoir. Then, magma pressure increased step by step with the rate of 0.5 Mpa. The aspect ratio(b/a) of magma body were taken as 1, 5 and 10. The results of calculation show that fractures start at the top of magma body of higher aspect ratios (that is dike-like magma body) and extend upwards forming fracture zone of funnel shape (Fig 2). The calculated vertical displacement of earth's surface indicates that a simple dome uplift is formed over a circular magma body, but that a central trough-like depression with flank uplifts is formed over magma bodies of higher aspect ratios(Fig.3)

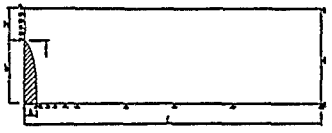


Fig.1 Model for calculation of magma intrusion.

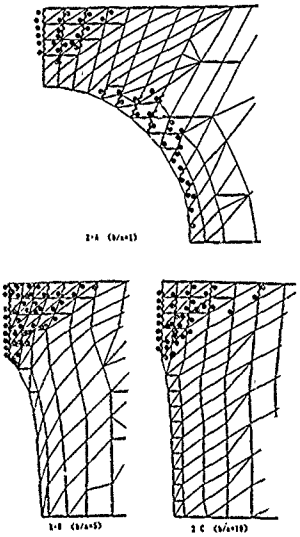


Fig 2 Extension of fractures(h=100m,b=200m) around the magma intrusion. Waved lines indicate tensile fractures and thick lines denote shear fractures. The numbers in circle indicate steps when fractures are formed.

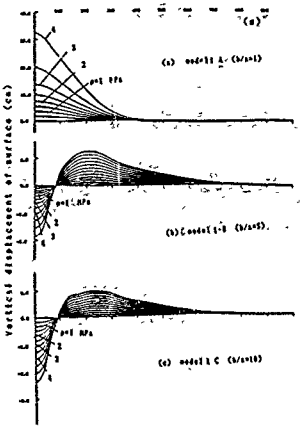


Fig.3 Vertical displacement of surface over subsurface magma intrusion(h=100m,b=200m).

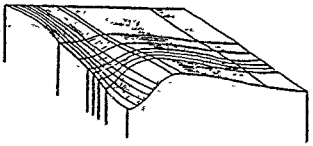


Fig.4 Sketch of vertical displacement of volcanic surface due to a subsurface dike-like intrusion calculated from the three dimensional RBSM model.

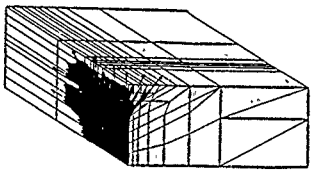


Fig.5 Extension of fractures around a dike-like magma intrusion estimated from the three dimensional RBSM model.

III TROUGH-LIKE SUBSIDENCE AS INDICATOR OF DIKE-LIKE MAGMA INTRUSION

In the Izu-Oshima volcano, trough-like subsidence during the time span which includes the eruption event, is clearly shown from the comparison of leveling data. The result of calculation indicates that the width of trough-like depression is nearly twice of the depth of top of magma intrusion(Fig.4). As the width of trough in the Izu-Oshima is about 4 Km in the narrowest part, the top of magma intrusion is expected at the depth of about 2 Km in the south-eastern part of Izu-Oshima volcano(Fig.5). The maximum subsidence of about 30 cm is expected by the magma pressure of about 6 MPa(Fig.6).

The results of three-dimensional calculation, also, clearly indicate that the trough-like subsidence in the Izu-Oshima volcano is a kind of "keystone graben" where the trough depression is formed by splitting of rock mass by the intrusion of dike-like magma wedge(Fig.4,5). Therefore, the subsidence of summit of Izu-Oshima volcano indicated the surface intrusion of magma and increase of magma pressure, that is, the subsidence is also the precursory phenomena of volcanic eruption in some cases.

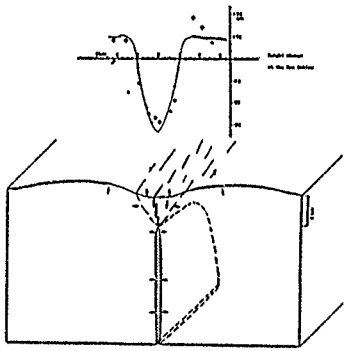


Fig.6 Model of graben subsidence over a large dike-like magma intrusion in the Izu-Oshima volcano. The upper diagram shows the vertical displacement measured in the Izu-Oshima volcano with the theoretical curve where the magma pressure is 6MPa.

IV CONCLUSION

Numerical analysis by discrete models is used successfully to simulate the formation of Keystone graben by subsurface intrusion of dike-like magma body beneath the Izu-Oshima volcano, Japan. Measurement of areal pattern of ground movement of volcanic surface is effective for the prediction of volcanic eruption, only with adequate numerical model for magma intrusion.

REFERENCES

Kawai,T.(1980):Some consideration on the finite element method,Int.J.Numerical Methods in Engineering, 16, 81-119.

Koide,H. and S.Bhattacharji (1975).Mechanistic interpretation of rift valley formation, Science,189,791-793.

Tada,T and M.Hashimoto(1987).The 1986 eruption of Izu-Oshima and crustal deformations, The Earth Monthly,9,396-403.

APPLICATION OF DISCRETE LIMIT ANALYSIS TO REINFORCED CONCRETE BEAMS

Tadahiko KAWAI, Science Univ. of Tokyo
Masatoshi UEDA, Takenaka Co.
Norio TAKEUCHI, Meisei Univ.
Harunori HIGUCHI, Abe-Kogyo Co.
Hiroaki KITO, Osaka City Univ.

ABSTRACT

Reinforced concrete beams subject to shear fracture were analyzed by a new discrete limit analysis method. A comparison was made between the results for the failure mechanism, failure process and maximum strength obtained through analysis by two types of incremental load methods.

1. INTRODUCTION

A family of new discrete models were proposed by Prof. Kawai in 1977 on the basis of the experimental evidence for solids under the ultimate state of loading [1]. These models, named RBSM, consist of rigid bodies and two types of connection springs, one of which resists dilatational deformation, while the other resists shear deformation.

The failure mechanism of reinforced concrete structures is complicated and it is extremely difficult to analyze their behaviour using existing analysis methods, which are usually based on the continuum mechanics approach. The authors have been engaged in the analysis of various types of reinforced concrete shear walls using RBSM's of reinforced concrete elements[2,3,4].

Modifications are required in the analysis algorithm for the incremental load method in analyzing deep beams, for example, in which the shear cracks in the concrete form severe fracture mechanisms. There are two types of algorithms for the incremental load method. One is Macal's method, in which the unbalanced forces are redistributed during the iteration, and the other is Yamada's method, in which the yielding, failure and unloading are judged for each integration point. In this paper, a proposal is made for a new incremental load algorithm, which is that of Yamada's method modified to accommodate cracking and recontact, and the results of the analysis for a deep beam using the constitutive relation for reinforced concrete obtained with the previously developed RBSM's are compared with test results to show its validity.

2. STRESS-STRAIN RELATION OF CONCRETE

The stress-strain relationships of concrete in uniaxial compression are approximated by a trilinear curve as shown in Fig.1. The stresses at the first and second yielding levels are Fc/2 and 0.95Fc, respectively. The tension-stiffening effect is also taken into consideration.

The failure surface of concrete loaded in its triangular plane is divided into 7 regions as shown in Fig.2 according to the states of stress. Stages 2 and 10 are decided according to the normal strain εn. Shear slip occurs when the stress state reaches stage 5 and the shear strain between the element boundaries is assumed to reach the constant value γu, and the spring value of the boundary is then replaced by zero at the stage of shear fracture.

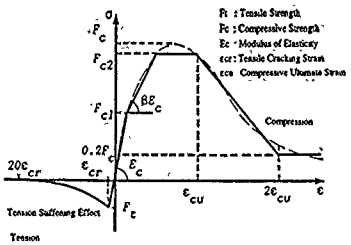


Fig. 1 Stress-Strain Relationship of Concrete

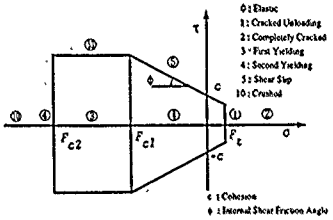


Fig. 2 Yield Failure Surface

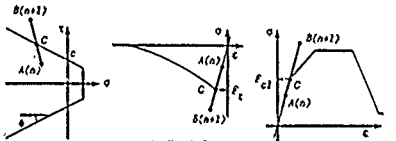


Fig. 3 Calculation of Incremental Load

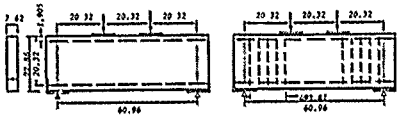


Fig. 4 Reinforced Concrete Beam Used in Analysis (Unit: cm)

Table 1 Concrete Material Constants (Units. kgf, cm)

Test Piece	Fc	Ft	c	φ	Ec	poisson's ratio	thickness
Model-1	237.0	23.6	32.568	37.0	$2.1*10^5$	0.167	7.62
Model-2	203.0	20.3	28.014	37.0	$2.1*10^5$	0.167	7.62

Table 2 Reinforcement Material Constant (Units. kgf, cm)

Type of Reinforcement	Modulus of Elasticity Es	Yield Strength Fy	Equivalent Thickness ts
Tensile Reinforcement	$2.1*10^6$	33250	4.072
Compressive Reinforcement	$2.1*10^6$	36200	2.243
Shear Reinforcement	$2.1*10^6$	22500	$8.306*10^{-1}$

3. ANALYSIS ALGORITHM

When stress is released by cracking and crashing, the force released may in turn cause cracking, compression failure and slips. Here a proposal is made for an algorithm, which is a modified version of Yamada's incremental load method and which can add the released force to the remaining load while counting the load and can simultaneously take account of slip failure, cracking and compression failure. As shown in Fig. 3, the minimum load increase $\bar{r}$ corresponding to various failure conditions at a given load increase P is obtained and the released force F is added to the remaining load. The remaining load at step n is expressed by the following equation.

$$P^{(n)} = \prod_{i=0}^{n-1}(1-r_i)P + \sum_{k=1}^{n}\left(\prod_{i=k}^{n-1}(1-r_i)F^{(k-1)}\right) \quad (1)$$

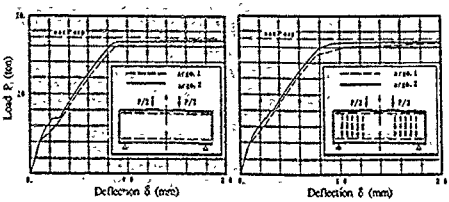
The process is repeated until the given load increase P and the released force are all used up.

4. NUMERICAL EXAMPLE

The deep beam test piece models, Models 1 and 2, with and without stirrups were analyzed by two types of incremental load methods, the normal Macal's method (Algo. 1) and the method proposed here (Algo. 2). The material constants used in the analysis are given in Tables 1 and 2. The load-displacement relationships for Models 1 and 2 are shown in Fig. 5. It can be seen from the figure that the differences between the incremental load methods originate in the cracking zone and the proposed algorithm (Algo. 2) gives solutions that correspond well to the failure load $_{max}P_{exp}$ obtained in the test. The deformation mode of Model 1 around the time of collapse indicate that the results obtained with Algo. 1 diverge from the test results near the centre of the span (Fig. 6). The failure characteristics at the time of collapse given by Algo. 2 are shown in Fig. 7. The formation of the resistance mechanisms in the deep beams is well represented in both Models 1 and 2.

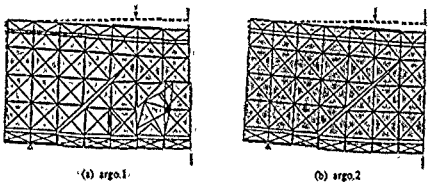
5. CONCLUSION

It was confirmed through analyses of the behaviour of reinforced concrete beams using RBSM's that they are capable of closely representing maximum shear strengths and failure characteristics of reinforced concrete structures. In particular, when combined with the analysis algorithm proposed here, it was made possible to carry out satisfactory evaluation of the conditions surrounding the progress of failure such as the occurrence of released force due to cracking and recontact.



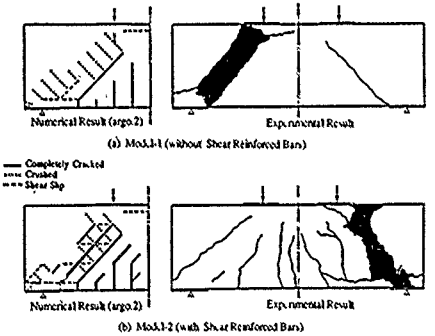
(a) Model-1 (without Shear Reinforced Bars) (b) Model-2 (with Shear Reinforced Bars)

Fig. 5 Relation between Working Load and Deflection at Span Centre



(a) argo.1 (b) argo.2

Fig. 6 Deformation Mode of Model 1 around Time of Collapse (P = 16 tons, 1/2 Zone)



(a) Model-1 (without Shear Reinforced Bars)

(b) Model-2 (with Shear Reinforced Bars)

Fig. 7 Failure Modes

1. Kawai,T., "New Discrete Structural Models and Generalization of the Method of Limit analysis",Finite Elements in Nonlinear Mechanics,Vol.2, Norwegian Institute of Technology, Trondheim, pp 885-906, 1977
2. Ueda,M., Takeuchi,N. and Kawai,T., "A Discrete Limit Analysis of Reinforced Concrete Structures", Proceeding of the International Conference on Computer Aided Analysis and Design of Concrete Structures Part II, pp.1369-1384, Sept 1984
3. Ueda,M. and Kawai,T., "Evaluation of a Discrete Limit Analysis to the Reinforced Concrete Shear Walls", Seminar on Finite Element Analysis of Reinforced Concrete Structures, JCI, pp.287-294, May 1985
4. Kei,T., Ueda,M and Kawai,T., "Nonlinear Analysis of a Reinforced Concrete Shear Wall by RBSM", Transaction of J CI, pp 449-456, Vol.7, 1985

APPLICATION OF THE RIGID BODIES-SPRING MODELS TO THREE DIMENSIONAL FRACTURE MECHANICS

Atsushi KIKUCHI*, Tadahiko KAWAI** and Noriyuki SUZUKI*

*Plant Engineering & Technology Bureau, Nippon Steel Corporation, Otemachi 1, Tokyo, Japan
**1st Faculty of Engineering, Science University of Tokyo, Shinjuku-ku, Tokyo, Japan

Abstract-The Rigid Bodies-Spring Models (abbreviated as the RBSM hereafter) are applied to the phenomenon of crack growth in the three-dimensional field and demonstrated the effectiveness of the models. The potential fracture criteria for the stable and unstable crack growth are discussed. Then numerical examples for each fracture problem are presented.

I. INTRODUCTION

A family of new discrete models named as the RBSM, which was proposed by one of the authors Kawai in 1976 [1], has been proved to be effective and suitable especially for analyses of nonlinear problems with large plastic deformation. With respect to the constitutive law of the connection springs, it is believed that the Mohr-Coulomb's law is the most suitable criterion for any material if the size of the elements is taken reasonably small according to accumulated results of numerical analysis.

In this paper, the practical application of the three-dimensional RBSM based on the previous work [2] is presented, and the potential fracture criteria are also discussed. The stable crack growth in an arbitrarily shaped initial crack and the unstable crack propagation of brittle material are simulated using these criteria.

II. OUTLINE OF THE RBSM

The detail of the RBSM has been already described in the previous paper [1]. The outline of the theory for the three-dimensional problems is as follows:

For the 3D problems the tetrahedral rigid element is used. Each element has 6 degrees of freedom at its centroid, which are assumed to be infinitesimal. For the analysis of crack growth, the tensile normal stress or strain is employed as the criterion parameter.

III. NUMERICAL RESULTS

A. Constitutive law and plastic region

In order to achieve an appropriate constitutive low for the RBSM, the numerical and experimental tests of the standard compact test specimen were carried out. The calculated plastic region around the crack-tip with each yield condition, the Von Mises' type or the Mohr-Coulomb's one, was indicated in Fig.1. As far as the calculated plastic zone is concerned, the Mohr-Coulomb's type shows fairly good agreement with the result [3][4] analyzed by the FEM and is considered more suitable for this kind of problem.

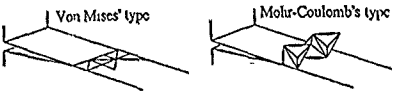


Fig.1 Calculated plastic region of crack tip

B. Criterion of stable crack growth

The criterion of the stable crack growth in ductile material was obtained by comparing analytical results for the previous test specimen with experimental ones. First, the load at which the crack-tip started propagating in the experiment (a big inflection point on the experimental COD-curve) is denoted by Pc, then the maximum strain around the crack-tip in the RBSM analysis for the specimen at the load Pc is defined as the criterion ε_c of the stable crack growth (Fig.2).

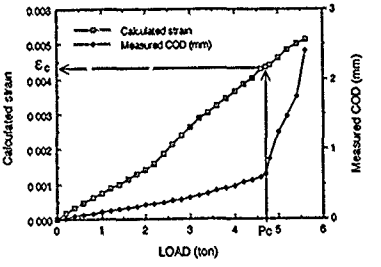


Fig.2 Criterion of stable crack growth

C. Simulation of V-shaped Crack Growth

In order to verify the generality of the constitutive law and the criterion, a V-shaped initial cracked body with the tensile loading was analyzed. The crack growth was simulated in the elasto-plastic analysis by cutting the springs of which the tensile strain ε_n reached the fracture criterion ε_c derived in the last subsection. (Fig.3)

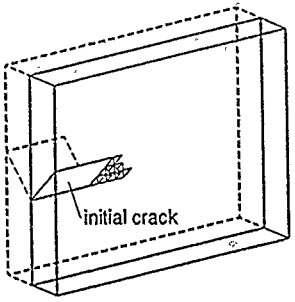


Fig.3 Calculated crack growth of V-shaped initial crack

D. Application to Dynamic Crack Propagation

A dynamic simulation of unstable crack propagation in the actual water vessel was carried out. At first the residual stress distribution of a welded plate was calculated by using the general purpose FEM program MARC and then obtained results were converted to the mesh subdivision for the RBSM-shell model [5]. Brittleness of the material was so strongly evident that the normal stress σ_n was employed for the criterion of the fracture and the critical stress value σ_c was determined in the same manner as the previous subsection.

Fig.4 shows the numerical results of the crack propagation. An initial crack existed along the right side of the welded rectangular plate and the spring at the initial crack tip was cut by force in the beginning of the transient analysis. The simulated route of crack propagation corresponded with the actual one.

IV. CONCLUSION

Aiming at the practical application of the RBSM the constitutive law for the three-dimensional crack problems is discussed. The calculated plastic zone of the crack-tip in our results indicate that the Mohr-Coulomb's law may be more suitable than the Von-Mises'. The cutting criterion for the springs of the RBSM is obtained by analyzing the compact test specimen itself, and the effectiveness of the criterion is verified in the application to an arbitrarily shaped crack. In an application to dynamic crack propagation of brittle material, the simulated crack propagation is in reasonable agreement with the actual crack propagated in the residual stress field of the welded plate in a water vessel.

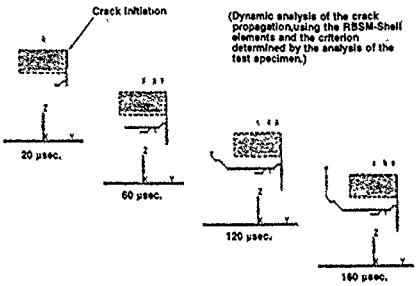


Fig.4 Simulation of the dynamic crack propagation

REFERENCES

[1]Kawai,T.:A New Discrete Model for Analysis of Solid Mechanics Problems, Journal of the Seisan Kenkyu, Institute of Industrial Science, University of Tokyo, Vol.29, No.4(1977), 208.

[2]Kikuchi,A. et al.:The Rigid Bodies-Spring Models and Their Applications to Three Dimensional Crack Problems, Comp. Mech. Appl. Mech. Engng. (to appear).

[3]Anderson,H.:A Finite-Element Representation of Stable Crack-Growth, J. Math. Phys. Solids, Vol.21(1973), 337.

[4]Miyamoto,H. and Miyoshi,T..Elastic-plastic Analysis of Crack Propagation, the 50th Annual Meeting, Proceedings of JSME. No.730-2(1973), 179 (in Japanese).

[5]Toi,Y. and Kawai,T.:Discrete Limit Analysis of Shell Structures (Part 4; Finite Deformation Analysis of Thick-Walled Shells), Seisan-Kenkyu, Vol.34, No.8(1982), 19 (in Japanese).

The Limit Analysis of Tapered Structural Members and Frames by a Discrete Model

NOGAMI Kuniei
Department of Civil Engineering
Tokyo Metropolitan University
Hachioji-city, Tokyo, 192-03, Japan

Abstract-This research aimed at the development of a discrete model for limit analysis of thin-walled members and frames. This model, having both geometrical and material nonlinearity, can deal with large deformation analysis. As a result, it was shown that the new discrete model was able to easily determine the development of a plastic region in a cross section in a load carrying problem, and it was able to consistently trace the equilibrium path considering from the stage of elastic deformation to plastic collapse by numerical example.

I. INTRODUCTION

Recently, many different procedures have been widely used for the discrete analysis of a continuum, but finite element method has been accepted as the most powerful analysis method available.

On the other hand, there is the rigid-body spring model as the limit analysis proposed from the standpoint of a discontinuum[1)2)]. The RBSM assumes the structural system to comprise rigid bodies and springs, and enables easy formulatation the geometrical and material nonlinearity and determination of the development of a plastic region.

This paper proposes a beam-column model which consists of the rigid-body spring system. The ultimate strengths of the tapered steel members and frames have been calculated by this model[3)~5)].

II. NEW DISCRETE MODEL

In a rigid-body spring model, a member is divided into a finite number of elements, and the elements themselves are assumed to be rigid bodies. The neighboring elements are connected with a spring system for resisting the relative motion on the boundary planes; in addition, mechanical characteristics are given to this spring system.

In this report, the author proposes the rigid-bar element model, as shown in Fig.1. This model utilizes a spring system composed of axial springs as a distributed spring system, and the torsion spring of St. Venant and shear springs, arranged in horizontal and vertical directions, as a concentrated spring system. In order to take warping deformation into consideration, the assumption of rigid elements was partially rejected, and unit warping ω was calculated as a continuum.

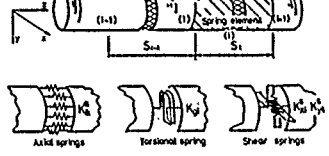


Fig.1 Rigid-bar element model

The determination of the spring constants should be carried out using the energy principle, and it is suggested that they be selected so that the virtual work $\delta\Pi_i$ by the spring system given in Fig.1 becomes

$$\delta\Pi_i = \delta\Pi_e \qquad (1)$$

for the same state of deformation, where $\delta\Pi_e$ is the virtual work as a continuum.

When a member is in an elastic state, it is assumed that the k-th axial spring takes charge of the partial cross-sectional area A_{ik} and is attached to the center of gravity of the grid. At this time, the axial spring constant is set according to

$$K^a_{ik} = EA_{ik}/S_i , \qquad (2)$$

where E is the modulus of elasticity and S_i is the length of spring element i.

The torsional spring constant and the shear spring constant are given as follows respectively.

$$\left.\begin{aligned} K_{\theta i} &= GJ/S_i \\ K^s_{xi} = K^s_{yi} &= GA/S_i \end{aligned}\right\} \qquad (3)$$

Where G is the elastic shear modulus and J is the torsion constant.

For the treatment of this model in an inelastic region, the following assumptions are introduced:

a) The yielding of materials is to be determined only by axial stress; the effect of shearing stress is neglected.

b) The stress-strain relationship between a hysteresis type body and a perfectly elasto-plastic body is used.

c) An linearized residual stress distribution is used.

d) The bending and bending-torsional rigidity must conform to the tangent modulus theory. Also, the torsional rigidity of St. Venant must conform to the plastic flow theory.

III. LIMIT ANALYSIS OF TAPERED BEAM-COLUMNS

When solving problems dealing with a tapered structural member, generally, the member is divided into several elements and treated as a collection of elements of uniform cross sections. Then the member is approximated by elements whose cross sections change in the stepwise manner.

RBSM can basically be treated in the same manner as the procedure explained above. As shown in Fig.2, since the mechanical characteristics of the cross section of each element of the tapered member are replaced with the spring constant of each "spring element", the elements of tapered members can be considered to be equivalent to the elements of uniform cross section found between the centers of gravity of adjacent rigid elements inclusive of the spring systems.

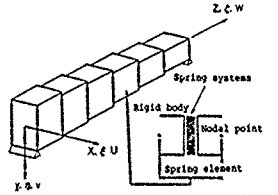


Fig.2 Spring element of a tapered member

As the method of changing the cross section, a tapered member whose web changes linearly in the direction of the length of the member was evaluated. As a variable parameter,

the web height ratio γ, the ratio of the web height at the minimum cross section to that at the maximum cross section, was used.

As an example of practical calculation, the displacement relationship of a simply supported beam-column with a rectangular cross section subjected to monoaxial bending. Here, $\gamma=0.5$, $M=0.2M_p$ (plastic moment), and residual compressive stress distribution $\sigma_{rc}=0.4\sigma_y$ were used. The number of elements divided was 20. Under these conditions, the results shown in Fig.3 were obtained.

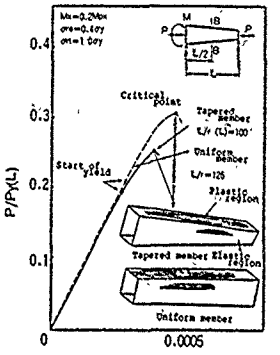


Fig.3 Deformation behaviour of tapered beam-column

The dotted line in the figuer shows the results for the member whose cross section was the same as the cross section of the central segment of the tapered member. Compared with the tapered member, a high critical load was obtained with the uniform beam-column. This can be explained by the fact that nonelastic behavior develops earlier in the tapered member than it does in the uniform member.

IV. LIMIT ANALYSIS OF FRAMES

In the case of the rigid body spring model, first, the model is constructed by assuming plastic hinges which can accomodate the possible range of plastic yield of the structure and by providing mechanical characteristics to the spring system at the hinges.

As a practical exampl., a problem of the ultimate strength in the direction of the span of the main tower of a suspension bridge is presented. The main tower receives horizontal reactive force due to the horizontal displacement via the main cable at the top of the tower; a model such as that shown in Fig.4 can be constructed.

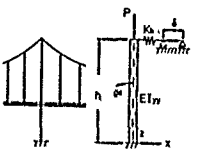


Fig.4 Tower model of suspension bridge

The cable can be modeled using an axial spring whose spring constant is $K_a=EI_{yy}/h^3\gamma_a$ ($\gamma_a=0.002$). The object of the analysis is the main tower, in which the relationship between the axial force at the top of the bridge, P, and the forced displacement, δ, is given by the following equation considering the entire suspension bridge.

$$P/P_y=421.4(\delta/h)^2+2.752(\delta/h)+0.295 \quad (4)$$

The tower consists of a monobox-type member (115 ×170 ×6mm) with h=2.8m. the yield load of $P_y=101.74t^{4)}$.

The critical load calculated by the present model, the experimental result and the value obtained by the difference method are compared in Fig.5. The F-δ curve obtained by the model agrees well with that obtained by difference method; they are represented by a single line. The horizontal reactive force of the cable obtained using the model in terms of absolute values was smaller than that obtained experimentally. The tendencies of the behavior of the F-δ relationship as determined by calculation and theory were essentially identical. Also, the critical axial force, P_{cr}/P_y, calculated using the model essentially agreed with that of the experimental result.

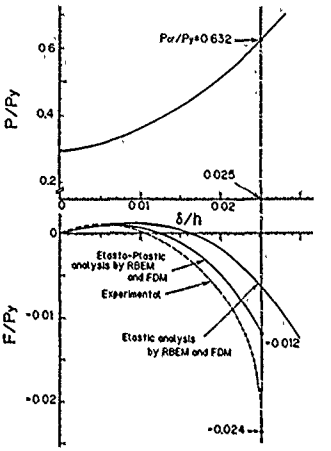


Fig.5 P-F-δ curves

V. CONCLUSIONS

The results of theso analyses are summerized as follows.

(1) When spring constants are determined by applying the variation principle, there is a possibility of the spring system model becoming too stiff or too flexible. However, the rigid-bar element model can uniquely select spring constants.

(2)For ultimate strength problems of tapered structural members and frames, the development of a yielded region in a cross section can be easily dealt with; in addition, a numerical solution having sufficient accuracy is obtained.

REFERENCES

1) Kawai,T.: Recent topic for a discrete limit analysis, Seisan seminar text, Course 116, Inst. of Ind. Science, Univ. of Tokyo, 1986.

2) Kawai,T.,Toi,Y. and Takeuchi,N.: Development of discontinuum mechanics(I-7), Journal of Seisan Kenkyu, 1980-1984.

3) Nogami,K. and Itoh,F.: Ultimate strength analysis of beam-column by finite rigid element model, Memoirs of the Faculty of Tech. Tokyo Metro. Univ., No.35, 1985.

4) Itoh,F., Nogami,K. and Ozaki,H.: In-plane buckling analysis of lib arch by rigid bar element model, Journal of Struct. Eng., Vol.33A, 1987

5) Nogami,K.. Formulation of a Rigid-Bar Element Model and its Application to Thin-walled Steel Members, Memoirs of the Faculty of Tech. Tokyo Metro. Univ., No.38, 1988.

6) Fukumoto,Y. and Omori,K.: Elastic-Plastic Behavior of Suspension Bridge Towers, Proceedings of JSCE, No.224, 1974.

ON A LIMIT ANALYSIS OF SOIL AND ROCK FOUNDATIONS BY MEANS OF THE DISCRETE ELEMENT MODEL

Tadahiko KAWAI
Department of Electrical Engineering
Science University of Tokyo
Tokyo 162, Japan

AND

Norio TAKEUCHI
Department of Civil Engineering
Teisei University
Tokyo 191, Japan

Abstract- Basing on the experimental evidence of solids under the ultimate state of loading, KAWAI proposed a family of new discrete models in 1976. These models consist of rigid bodies and two type of connection springs, one of which resist the dilatational deformation, while the other shear deformation. In this paper, application of these models is proposed to soil and rock mechanics.

1.INTRODUCTION

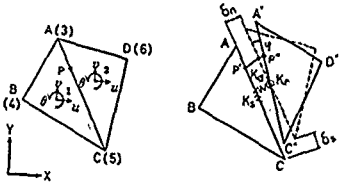
To consider the soil and rock materials as continua they are generally too nonuniform, inhomogeneous and easy to slip internally under applied loading. The finite element method treat the soil and rock masses on the continuous materials. On the other hand, the simplified method is a practical but rather crude method which has been developed by ingenious use of elasticity and plasticity theories because the stress field is determined without consideration on the displacement and strain hysteresis.

Considering such a status of existing methods KAWAI proposed a family of new discrete models. In these models structures or solids are idealized as a set of rigid elements interconnected by two types of spring system, one of which resist the dilatational, the other shearing deformation. Therefore sliding or separation of two adjacent elements can be made easily.

In case of the granular materials influence of crack initiation due to tensile load often can not be neglected. In view of this, present authors developed a new algorithm which may be applicable to analyze the coupled failure of solids due to slippage, tensile cracking and solids contact. In this paper application of these models to analysis of foundation structures will be attempted and the general method of discrete limit analysis will be described with some verification examples.

2.FORMULATION OF TWO DIMENSIONAL RBSM

For simplicity, consider two dimensional rigid triangular element of RBSM as shown in Fig.1. They are assumed to be equilibrium with external loads and reaction forces of the spring system which is distributed over contact surface of two adjacent bodies.



(a) before deformation (b) after deformation

Fig.1 Two dimensional rigid triangular element

Rigid displacement field is assumed in each element, whose nodal displacement are given by the displacement (u,v,θ) of the centroid as shown in Fig 1. Therefore, the relative displacement vector δ of the arbitrary point P can be derived as follows:

$$\delta = \sum_{i=1}^{2} B_i u_i \qquad (1)$$

$$u_1 = (u_1\ v_1\ \theta_1)^t,\ u_2 = (u_2\ v_2\ \theta_2)^t,\ \delta = (\delta_s\ \delta_n)^t$$

$$B_1 = \begin{bmatrix} -l_1 & -m_1 & l_1(y-y_1)-m_1(x-x_1) \\ -l_2 & -m_2 & l_2(y-y_1)-m_2(x-x_1) \end{bmatrix}$$

$$B_2 = \begin{bmatrix} l_1 & m_1 & -l_1(y-y_2)+m_1(x-x_2) \\ l_2 & m_2 & -l_2(y-y_2)+m_2(x-x_2) \end{bmatrix}$$

$l_1 = \cos(\overline{x},x)$, $l_2 = \cos(\overline{x},y)$, $m_1 = \cos(\overline{y},x)$, $m_2 = \cos(\overline{y},y)$

On the other hand, the following relation are obtained from the definition of the spring constants:

$$\sigma = D \cdot \delta \qquad (2)$$

$$\sigma = (\tau_{s\overline{x}}\ \sigma_n)^t,\quad D = \begin{bmatrix} k_{s\overline{x}} & 0 \\ 0 & k_n \end{bmatrix}$$

$$k_n = \frac{(1-\nu)E}{(1+\nu)(1-2\nu)h},\quad k_{s\overline{x}} = \frac{E}{(1+\nu)h}$$

where $h = h_1 + h_2$ is the projected length of a vector connecting centroids along the normal drawn, and $\tau_{s\overline{x}}$ and σ_n is tangential and normal stress respectively.

Basing on the above preliminaries, the strain energy expression of the in-plane element V can be obtained as the following:

$$V = \frac{1}{2}\int_{l_{ij}} (\delta^t \cdot D \cdot \delta) dS = \frac{1}{2} u_i^t \int_{l_{ij}} (B^t \cdot D \cdot B) dS u_i \qquad (3)$$

Applying Castigliano's theorem, the following stiffness equation can be derived:

$$P = \frac{\partial V}{\partial \delta} = K \cdot u \qquad (4)$$

$$P = (X_1, Y_1, M_1, X_2, Y_2, M_2),\ u = (u_1, v_1, \theta_1; u_2, v_2, \theta_2)$$

where K is a (6 × 6) symmetric matrix and P is a nodal load vector

3.CONSTITUTIVE LAW

In the RBSM, present authors considered that reaction stresses inducted are not tensor but vector, and consequently Coulomb's condition may be most realistic constitutiv law for such a discrete system representing granular materials. In case of the granular materials like soils, it is commonly observed that normal stress is relieved as soon as it reaches σ_t as shown in Fig 2.

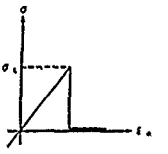


Fig.2 Stresss-strain relation at the tensile failure

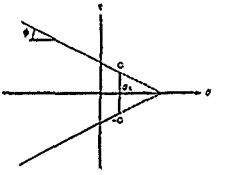


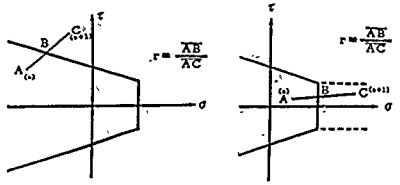
Fig.3 Modified Coulomb's condition

Consequently the yield condition of soil-like materials can be modified as shown in Fig.3.

For determination of spring constants in shear failure plastic flow rule is adopted.

4.PROPOSED ALGORITHM

A new algorithm is proposed by applying the incremental loading procedure developed by YAMADA. In Yamada's method, necessary rate of the load increment to yield the most heavily stressed element can be calculated by stress distribution and load increment at the present stage as shown in Fig.4. From this condition the required rate of load increment r can be calculated. Once the stress point lies on the failure curve, it may move according to the plastic flow rule until the unloading occurs.



(a) shear failure (b) tensile failure

Fig.4 Rate of load increment

Similar calculation must be model in case of the tensile failure as shown in Fig.4. All the possible rate of load increment corresponding to failure patterns should be calculated in all the elements and the minimum rate of load increment must be determined at each step as the rate of load increment corresponding to the next step.

Stress relaxation is usually follows by the tensile failure. If Yamada's method is applied to this stress relaxation process exactly, endless calculation should be repeated correspondingto the tensile failure which may occur continuously.

The load P^{i+1} at the (i+1)th step can be calculated by using load P^i and rate of load increment r_i at the present step (i) as follows:

$$P^{i+1} = (1 - r_i)P^i \qquad (5)$$

Therefore, in case of shearing failure, residual load P^n at the n_{th} step can be obtained by using initial load P as follows:

$$P^n = \prod_{i=0}^{n-1}[(1 - r_i)]P \quad (r_0 = 0) \qquad (6)$$

This is the same result with Yamada's method.

On the other hand, if stress relaxation will caused by crack initiation, relieved forces are taken into account as follows.

$$P^n = \prod_{i=0}^{n-1}[(1 - r_i)]P + \sum_{k=1}^{n}(\prod_{i=k}^{n}[(1 - r_i)]F^{k-1}) \qquad (7)$$

where F^k is the relieved force at k_{th} step.

Here r_{total} implies the cumulative rate of load increment and it can be defined as follows:

$$r_{total} = \sum_{k=1}^{n}(\prod_{i=0}^{k-1}[(1 - r_i)])r_k \qquad (8)$$

The calculation must be repeated until $r_{total} = 1$ in each stage of loading. Fig 5 shows outline of the flow diagram of the proposed algorithm.

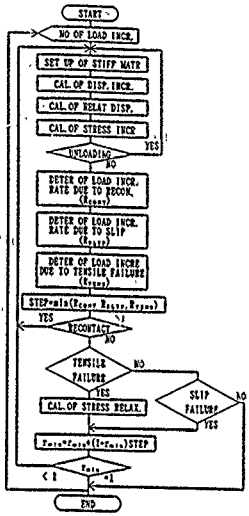


Fig.5 Flow diagram of the present algorithm

5.NUMERICAL EXAMPLE

Fig 6 shows the numerical model for a anker block and material constants used. In the present analysis effect of the gravitational load was neglected. The horizontal load at the anker block was applied in step by step mannnner taking the incremental as 10t, 10t, 5t, 5t and 5t. Fig7. shows the slip line pattern of the solution at the step 5. In this figure it can be seen that not only slip lines but also tensile cracking may spread on the front region of a given block and may be considerably different from that of the previous solution. the displacement mode corresponding to this step are shown in Fig.8. From this figure separation of the soil on the rear wall of block can be seen.

6.CONCLUSION

A new algorithm was developed by which coupled failure due to shear and tensile loads can be treated. Although a numerical exsample is very simple, it is believed that the present method may be useful to failure analysis of various soil and rock engineering problems.

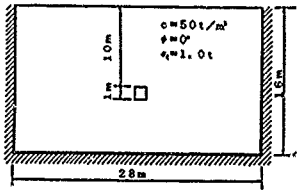


Fig.6 Numerical model and material constants

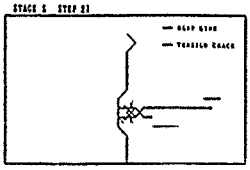


Fig.7 Slip line pateern

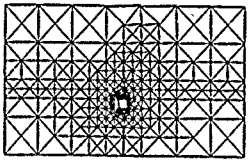


Fig 8 Displacement mode

ON THE SIMPLIFIED DISCRETE LIMIT ANALYSIS FOR SLOPE STABILITY ANALYSIS

Tadahiko Kawai
Professor
Science University of Tokyo
1-3,Kagurazaka,Shinzyuku,Tokyo 161,Japan
Norio Takeuchi
Associate Professor
Meisei University
2-1-1,Hodokubo,Hino,Tokyo 191,Japan

Kei Yada
Engineering Research Institute:Sato Kogyo Co.,Ltd.
47-3,Sanda,Atsugi,Kanagawa 243-02,Japan

Abstract-A numerical procedure for the slope stability analysis which combins the slice method in RBSM is shown in this paper.
Feature in the present method is as follows:
(1)Plastic work is evaluated only along arc.
(2)Tensile failure is also taken into account.
(3)Mesh division is unnecessary.
(4)It is designed to compute with data of slice method.

1.INTRODUCTION

Slice methods are often applied when slope stability problem is discussed. The Fellenius method is most popular among slice methods, but in this method the force between each piece of slice is not considered. Therefore the methods such as the Bishop, the Janbu, the Spencer were proposed and they are considered the effects of statically indeterminate force. Plastic condition, however, is not taken into account in these slice methods, because they are based on the limit equilibrium method.

Finite element method is also used to analyze slope stability problems, but there are some problems that definit slipe line may not be obtained clearly and making input data is laborious work comparing with slice methods.

On the other hand, RBSM (Rigid Bodies - Spring Model) has been proposed as a physical model especially suitable for limit analysis of solids and by which mutual slip movement of two adjacent element can be simulated. It's usefulness has been duly verified by solving many collape probrems where slippage is considered problems. The solution is influenced by the mesh division and the upper bound solution is obtained because of this specific character.

A numerical procedure for the slope stability analysis which combins the slice method in RBSM is shown in this paper. Feature in the present method is as follows:

(1) Plastic work is evaluated only along a circular arc.
(2) Tensile failure is also taken into account
(3) Mesh division is unnecessary.
(4) It is designed to compute with data of slice method.

2.ALGORITHM OF PROPOSED METHOD

To simplify the explanation of the proposed method, the case which slipe lines are modeled by circular arc is taken as an example.

The procedure of analysis is as follows:

(1) As shown in Fig.1, the region which is defined by shape of slope and slipe line is devided into slice. This process is same as the ordinary slice method.
(2) Integral points(○) are assumed between circle points(●).
(3) As shown in Fig.2, the degrees of freedom are rigid body displacements (u,v,θ) at the center of gravity of each slice element. Then stiffness matrix of each slice element is obtained.
(4) Driving forces are calculated from the weight of each slice element.
(5) As shown in Fig.2, surface stresses (σ,τ) at each integral point are computed by the descrete limit analysis.
(6) The safty factor is computed by the surface stresses.

The safty factor can be given by the following equation:

$$FS=\frac{\Sigma\,(c+\sigma\tan\phi)\,\ell}{\Sigma\,\tau\,\ell}$$

where: c is cohesion
ϕ is internal frictional angle
σ is normal stress
τ is shear stress
ℓ is length of slip lines

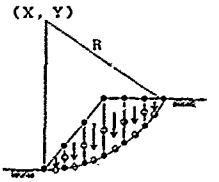


Fig.1 Model of slope

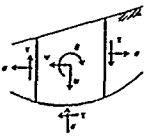


Fig.2 Slice element

3.TECHNIQUES OF NON-LINEAR ANALYSIS

Two techniques for solving non-linear analysis are discussed in this paper. One is the stress transfer method in which the tensile failure can be taken into consideration easily. Another is the Yamada's method in which the yielding failure and unloading are judged exactly. In the stress transfer method the corrective increment stress is caluculated only by the normal stress as shown in Fig.3. Therefore the stress paths of both will be different.

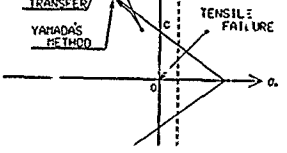


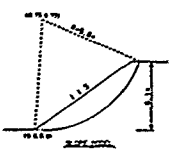
Fig.3 Representation of stress path

4. NUMERICAL EXAMPLES

4.1 COMPARISON OF SAFTY FACTOR WITH THE ORDINARY METHODS

A test model is shown in Fig.4. In this case, the center and the radius of a circular arc are fixed. The safty factors compared with the ordinary slice methods are shown in Tab.1. The results obtained from this table are summarized as follows:

(1) The safty factor which is computed by elastic analysis is smaller than the Fellenius method.
(2) The safty factor which is computed by the stress transfer method is almost the same as the Yamada's method.
(3) The safty factors which is computed by the stress transfer method and the Yamada's method are situated between the Fellenius and the Bishop.
(4) In the case of the stress transfermethod, The safty factors which is computed by plastic within a tensile strength differs little from plastic.



γ	2.0 t/m³
C	0.499 t/m²
φ	32 °
σt	0.1 t/m²

Fig.4 Test model

Tab.1 Safty factor

FELLENIUS	1.43
BISHOP	1.76
JANBU (M/N)	1.63
ELASTIC	1.408
YAMADA'S METHOD	1.456
STRESS TRANSFER (ELASTIC)	1.451
STRESS TRANSFER (ELASTIC WITH TENSILE STRENGTH)	1.453

4.2 STATICALLY INDETERMINATE FORCES

Statically indeterminate forces are shown in Fig.5. In elastic analysis, horizontal forces E become large tension near the top of slope, but in plastic analysis the tensile forces are decreased owing to occurrence of slip lines. Besides in elastic analysis vertical forces T which indicate shear forces are greatly changed from toe to top of a slope, but in plastic analysis the range of change becomes small. Then the local safty factors computed by plastic analysis are lower than those by elastic analysis except the region of slippage. In the region of slippage, local safty factors are 1.0 in plastic analysis. The stress transfer method and the Yamada's method have similar distribution of statically indeterminate forces.

4.3 THE STRESS TRANSFER METHOD WITHIN A TANSILE STRENGTH

Fig.6 shows the results of two methods. One is the method which is taken no account with a tensile strength with a broken line. Another is the method which is taken into account of a tensile strength with a solid line. Taking account of a tensile strength, the tensile region which is occurred at a top of slope is disappeared. Then the length of slipo line becomes longer. Both horizontal forces E and vertical forces T are replaced towords the compressive side.

4.4 PIGION OF TENSILE FAILURE USING STRESS TRANSFER METHOD

Hoek's study shows that in the case of slope having a face angle of 30 degrees in a drained soil with a friction angle of 20 degrees, the circle center of the critical failure is located at (0.2H,1.85H) and the critical tensile crack is at a distance 0.1H behind the top of the slope. The result of H=10m using the stress transfer method within a tensile strength is shown in Fig.7. As shown in this figure, the tensile failure is located at a distance 0.68m(0.07H) and slipe lines are occurred from the bottom of tensile failure. Because of the size of the breadth for each slice element, the location of tensile failure is reasonable.

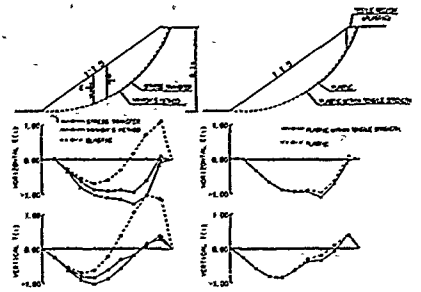


Fig.5 The stress transfer method and the Yamada's method

Fig.6 The stress transfer method within a tensile strength

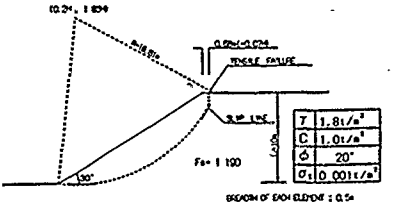


Fig.7 Tensile failure

5.CONCLUSION

A numerical procedure for the slope stability analysis which combins the slice method in RBSM was proposed and applied it to some numerical examplos.

The principal results and conclusion of the present study are:

(1) The numerical data for the ordinary slice methods are available to this method.
(2) The safty factor can be obtained fairly easily and the value is reasonable, compared with the ordinary slice methods such as the Bishop, the Janbu and so on.
(3) The results of the stress transfer method and the Yamada's method have a littie difference.
(4) In the case of including tensile failure, the stress transfer method within a tensile strength is useful.

IMPACT FRACTURE ANALYSIS OF REINFORCED CONCRETE BEAMS BY RIGID BODY SPRING MODEL

KEIICHIRO SOHODA[1], HIROAKI KITOH[1] and ATSUSHI KAMBAYASHI[2]

[1]Department of Civil Engineering
Osaka City University
Sumiyoshi, Osaka 558 JAPAN

[2]Technical Research Laboratory
Takenaka Komuten Co., Ltd.
Koto, Tokyo 136 JAPAN

Abstract - This paper deals with an explicit finite difference scheme with rigid body spring model for some impact problems on elastic plane beams and reinforced concrete beams. The beams are divided into a finite number of small rigid bodies, which are connected with springs distributed the contact area of neighboring bodies. Equations of motion on each rigid body are expressed in a finite difference form on time and then these equations are solved both numerically in an explicit form. Present numerical results show good agreement with the exact solution for an elastic plane beam and experimental results for a RC beam.

I. RIGID BODY SPRING MODEL

Considering two triangular rigid plate elements i,j which are connected by two different types of springs K_n and K_s at two evaluation points on contact area as shown in Fig.1(a,b), stiffness of the springs is expressed from the plane stress condition as follows:

$$K_n = \frac{\ell_{ij}\, E}{2h_{ij}\,(1-\nu)} \quad , \quad K_s = \frac{\ell_{ij}\, E}{2h_{ij}\,(1+2\nu)} \tag{1}$$

Where E and ν are Young's modulus and Poisson's ratio, and other symbols should refer to Fig.1. Centroidal displacements of each element are denoted by (u_i, v_i, θ_i) and (u_j, v_j, θ_j) respectively. Relative displacements on the evaluation points are given by the following equations(see Fig.1(c)):

$$\delta_n = -(\delta_{ni} + \delta_{nj}) \quad , \quad \delta_s = -(\delta_{si} + \delta_{sj}) \tag{2}$$

where

$$\delta_{ni,j} = u_{i,j}\cos\alpha_{i,j} + v_{i,j}\sin\alpha_{i,j} - [(y_p - y_{gi,j})\cos\alpha_{i,j} - (x_p - x_{gi,j})\sin\alpha_{i,j}]\theta_{i,j}$$

$$\delta_{si,j} = -u_{i,j}\sin\alpha_{i,j} + v_{i,j}\cos\alpha_{i,j} + [(y_p - y_{gi,j})\sin\alpha_{i,j} + (x_p - x_{gi,j})\cos\alpha_{i,j}]\theta_{i,j}$$

where $x_{gi,j}$, $y_{gi,j}$ and x_p, y_p are coordinates of the centroid of plates and the evaluation point, respectively. Here rotational displacement $\theta_{i,j}$ are assumed to be very small. Spring forces interacted on the two elements are given by .

$$N_{ij} = K_n\,\delta_n \quad , \quad S_{ij} = K_s\,\delta_s \tag{3}$$

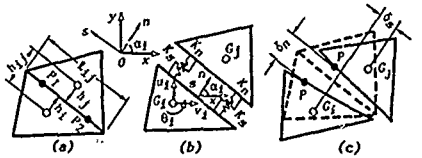


Fig.1 Rigid Body Spring Model

II. EXPLICIT FINITE DIFFERENCE SCHEME

At time t, acceleration of each element is subjected to the equation of motion,

$$\ddot{u}_{i,t} = \Sigma X_i/\varpi_i \quad , \quad \ddot{v}_{i,t} = \Sigma Y_i/\varpi_i \quad , \quad \ddot{\theta}_{i,t} = \Sigma M_i/I_i \tag{4}$$

where ΣX_i, ΣY_i, ΣM_i are summation of all the forces applied to element i, and ϖ_i and I_i are mass and moment of inertia of elements i, respectively.
Velocities and displacements at time $t+\Delta t$ are obtain by a direct time integration using forward finite difference expressions as follows:

$$\dot{u}_{i,t+\Delta t} = \dot{u}_{i,t} + \ddot{u}_{i,t}\,\Delta t \quad , \quad \dot{v}_{i,t+\Delta t} = \dot{v}_{i,t} + \ddot{v}_{i,t}\,\Delta t \quad , \quad \dot{\theta}_{i,t+\Delta t} = \dot{\theta}_{i,t} + \ddot{\theta}_{i,t}\,\Delta t \tag{5}$$

$$u_{i,t+\Delta t} = u_{i,t} + \dot{u}_{i,t+\Delta t}\,\Delta t, \quad v_{i,t+\Delta t} = v_{i,t} + \dot{v}_{i,t+\Delta t}\,\Delta t, \quad \theta_{i,t+\Delta t} = \theta_{i,t} + \dot{\theta}_{i,t+\Delta t}\,\Delta t \tag{6}$$

III. ELASTIC PLANE BEAM

First example concerns a problem of elastic stress wave propagations in a two dimensional plane elastic beam subjected to an impact step load. A rectangular beam whose two edges are simply supported is considered, and element mesh division and boundary element condition used here are shown in Fig.2. Figure 3 shows stress history of σ_y at the mid point of beam (x=0,y=h/2) as compared to the exact solutions by the two dimensional elasto dynamic theory. The numerical solution exhibits dispersion and spurious oscillations behind wave front However, wave arrival time and amplitudes are correct and the center of oscillations approaches the exact solution. In the figure, the abscissa indicates a non-dimensional time, $T = c_1 t/h$, where $c_1 = \sqrt{E/[\rho(1-\nu^2)]}$, ρ = density.

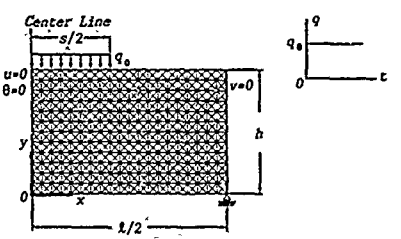


Fig.2 Mesh division

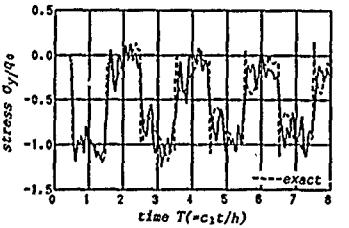


Fig.3 σ_y response at point(x=0, y=h/2)

IV. REINFORCED CONCRETE BEAM

Second example concerns an impact fracture problem of

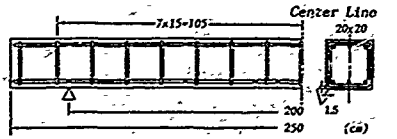


Fig.4 Reinforced concrete beam

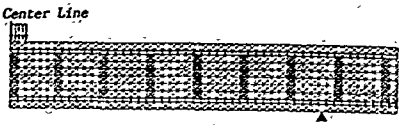


Fig.5 Element mesh division; dotted part shows the element including reinforced bar

reinforced concrete beam. Model considered here is the same as one of the test specimens in the experiments of reference[2]. The impact loading system used in the experiment was operated by use of compressed N_2 gas through a striking hammer consisting of a hard steel cylinder with diameter of 9.8cm. Impact velocity of 10.7m/s and weight of 70kgf. Figure 4 shows the test specimen and Fig.5 shows the element mesh division used for numerical analysis. Constitutive relation of concrete used in this analysis includes the effects of cracking and shear-slippage as given in Fig.6. And the main and shear reinforcing steel bars are elastic-perfectly plastic (yield stress is 3000kgf/cm^2). Figure 7 shows the two fracture modes obtained, namely Fig.7(a) is a bending fracture mode under a comparatively slow monotonously increasing load(10tf/ms), and Fig.7(b) is a shear and negative bending fracture mode under an impact load with the imposed hammer velocity, V=10.7m/s. The later mode is similar to the fracture modes observed in the experiments as shown in Fig.8. Figure 9 shows the comparison of strain histories of the top and bottom reinforcing bars. Fair agreement is seen between the numerical results and the experimental results at a point away from the span center. But it seems that considerably difference between them is yielded at the span center because of the effect of three dimensional local failure near the loading point.

V. CONCLUSION

The proposed method showed good applicability to impact fracture problems including crack or shear failure of reinforced concrete beams as well as stress wave propagation problems of elastic plane beams.

References

[1] Kawai,T.: J. Soc. Naval Arch. Japan, Vol.141, 1977.

[2] Ito,C. et al: Central Research Institute of Electric Power Industry of Japan, Rep. No.383046, 1984.

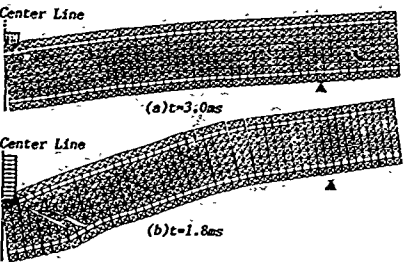


Fig.7 Fracture modes. (a)Slow-increasing load(P=10tf/ms) (b)Impact load(V=10.7m/s)

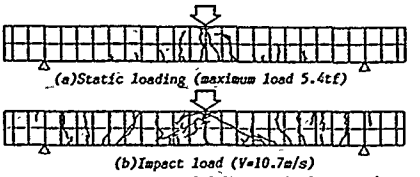


Fig.8 Cracking pattern and failure mode for test beams

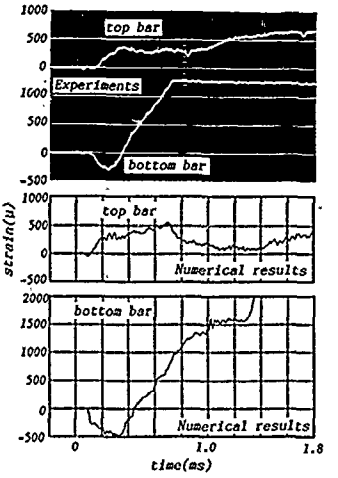


Fig.9 Comparison of strain response of re-bar

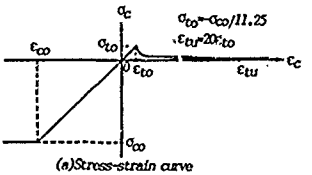
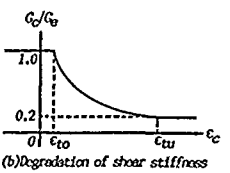
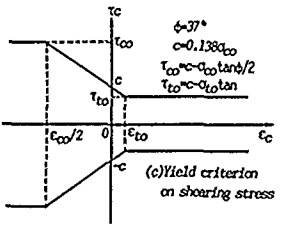


Fig.6 Constitutive relation of concrete





APPLICATIONS OF 3D RIGID-BODY SPRING MODEL TO STEEL AND CONCRETE COMPOSITE SLABS

HIROAKI KITOH and KEIICHIRO SONODA

Department of Civil Engineering
Osaka City University
Sumiyoshi, Osaka 558 JAPAN

Abstract - Steel and concrete-composite slabs, described herein, are two layered thick plates consisting of concrete and thin steel plate with mechanical shear connectors. Their behavior up to failure is simulated using 3-dimensional Rigid Body Spring Model considering material nonlinearity. Verification study can show the validity of the model, even when the coarse subdivision is used based only on major failure modes to control the slabs at the ultimate state of loading.

I. INTRODUCTION

The concept of Rigid Body Spring Model: RBSM[1] is to idealize structures macroscopically to the assemblage of rigid bodies together with springs, based on the failure mechanism observed. It has an advantage of comprehensive expression to the discontinuous phenomenon due to separation and slip to play a vital role at structural collapse.

Steel and concrete composite slabs[2,3] have excellent performance on load carrying capacity and execution workability in comparison with ordinary reinforced concrete slabs. Thus they have been applied to plate members such as bridge decks, building floors and the another in civil engineering field.

At the ultimate state of loading, the slabs exhibit some kinds of complicated and solid failure modes. Moreover, the modes change dependent upon shear connector arrangement, applied load condition and so on. We have, therefore, carried out 3-dimensional nonlinear analysis of the slabs using RBSM. It is significant to simulate their behavior up to failure, because it could be reflected on establishment of their rational design codes when the model can predict the behavior sufficiently.

II. MODELING OF CONSTITUTIVE RELATION

In RBSM formulation, material properties are introduced to normal and tangential springs on the interfaces of neighbor rigid body surfaces. According to the material properties, the couple of spring stiffness of the model are determined and can associate relative displacements of neighbor rigid body with interfacial forces. The nonlinear characteristics of the materials made the composite slabs up are modeled as follows:

A. Concrete Material

Figure 1(a) shows the relationship between normal stress and normal strain of concrete taking account both of cracking in tension and crushing in compression. The shear slip of concrete is subjected to the analogous failure surface[4] with Mohr-Coulomb's as shown in Fig. 1(b). Furthermore, within the failure surface, the elastic relation in shear is applied.

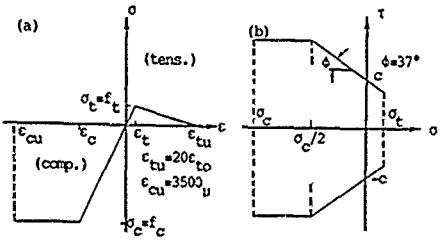


Fig.1 Concrete material model
(a)σ - ε relationship, (b)Failure surface

B. Steel Material

Steel plates are assumed to be elastic and perfectly plastic. Interaction between normal stress and shear one for the failure surface of concrete is ignored.

C. Shear Connector

The shearing-force - slip relationship on interface through an embodied stud connector in concrete is expressed as a function of its strength. Q_u[5] as shown in Fig.2. In addition, its longitudinal action is regarded as an elastic bar element.

The effect of the stud is, hence, evaluated by two modelings. One is to allocate a couple of springs with the above property to every point of stud and no interaction of concrete and steel plate except studpoints. The other is to spread the stud strength over appropriate area surrounding studs, and the relation between shearing stress and slip on the concrete - steel plate interface is given by an elastic perfectly plastic curve. The former and the latter are called discrete and smeared model, respectively.

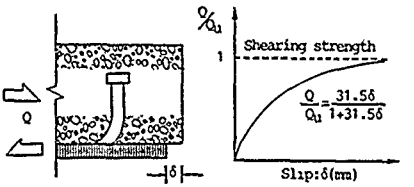


Fig.2 Mechanical behavior of stud connector[5]

III. VERIFICATION STUDY

A. Test Slabs

All of the slabs we carried out loading tests for verification were square slabs with sides 1600mm long, and were simply supported with 1375mm span length. Concrete of upper layer was 120mm high and steel plates of lower one was 6mm thick. Concrete was 396kgf/cm² in compressive strength: f_c and 28kgf/cm² in tensile one: f_t, and steel was 3574kgf/cm² in yielding point: f_{sy}. As shear connector, headed studs with 80mm height and 13mm diameter were welded on steel plates in the following two ways; In slab #1, they were arranged out of the supported edges at intervals of 250mm, while in slab #2, over the whole at those of 125mm. Square patch load with sides 125mm(Load #1) or 375mm(Load #2) long was applied at central portion on the top surface of the slabs.

The observed failure modes of the slabs could be classified into two types; One was bond slip failure due to distinct slip associated with cracking on all edge sections, shown in Fig.3(a) for the slab having coarser stud arrangement. The other was punching shear failure pushing out of concrete locally at the loading portion in Fig.3(b) for that having intimater one. Additionally, in both the modes, steel plate had not yielded.

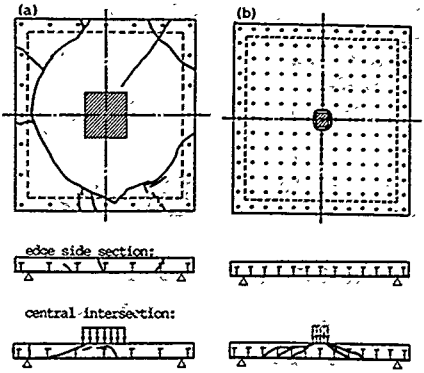


Fig.3 Test slabs with cracking of concrete observed
(a)Slab #1; Bond slip failure under Load #2
(b)Slab #2; Punching shear failure under Load #1

B. 3-dimensional RBSM Idealization

To give full play to the ability of RBSM, we intended, in particular, to subdivide 3-dimensionally the slab as coarse as possible considering only the major failure modes observed; Namely, it was to form 1) a corned shape division expanding from the sides of loading area corresponding to the punching shear crack, 2) a polygonal one near the supported edges to express the accompanying negative bending crack and 3) one on steel plate - concrete interface to express slip and separation, as shown in Fig.4 with referring to Fig.3. The slabs in 1/4 space dotted in Fig.4 were analyzed owing to symmetry.

A direct iteration scheme was employed modifying the secant modulus of stiffness at each step of computation, according to a current state on the nonlinear material models. The reference points to the current state coincided with the sampling points for numerical integration to evaluate the stiffness of the model.

C. Results

The load - deflection curves obtained numerically using the discrete model(Model #1) or the smeared model (Model #2) for stud connectors are shown in Fig.5 with that observed in the test. Model #1 predicted the failure loads sufficiently for slab #1 having coarser stud arrangement. On contrary, Model #2 gave good results for slab #2 having intimater one. Thus it was essential for the successive simulation to evaluate the effect of the stud adequately dependent on the arrangement on the modeling.

Figure 6 shows the failure modes obtained, those gave the failure loads in good agreement with the observed ones. The numerical results also expressed the satisfactory failure modes. One can find the occurrence of significant slip on the steel - concrete interface in Fig.6(a) characterizing the bond slip failure. Furthermore, the situation pushing out the corned region downward at loading area corresponding to the punching shear one can be found in Fig.6(b). In both the modes, steel member had not yielded similarly to the test results.

IV. CONCLUSION

Using 3-dimensional Rigid Body Spring Model, we have simulated the behavior of the steel and concrete composite slabs up to failure and examined the numerical results obtained in comparison with those observed in the loading tests both on the failure loads and modes.

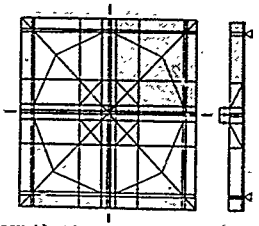


Fig.4 3D RBSM idealization which has 98 rigid bodies and 588 degrees of freedom in 1/4 space dotted

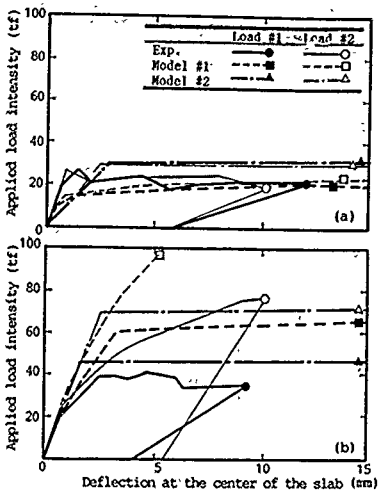


Fig.5 Load - Deflection curves. (a)Slab #1, (b)Slab #2; Numerical and experimental results

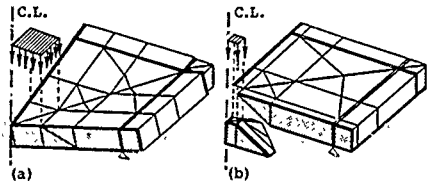


Fig.6 Failure modes (1/4 space); Numerical results
(a)Bond slip failure, (b)Punching shear failure

Thus we can conclude that the model can predict satisfactorily the mechanical behavior of the slabs even when the coarse subdivision is used based only on major failure modes, provided with the adequate modeling of the stud connectors dependent upon its arrangement.

References

[1] Kawai,T.: Proc. Soc. Naval Arch. Japan, No.141(1977)
[2] Sonoda,K. et al. Proc. Eng Found. Conf., ASCE(1988)
[3] Kitoh,H. et al: IABSE Sympo. Brussels(1990)
[4] Leda,M. et al. Int. Conf. Computer Aided Analysis and Design Concrete Struc., Yugoslavia(1984)
[5] Oilgaad,J.G. et al: Eng. J. AISC, Vol.18, 2(1971)

DISCRETE LIMIT ANALYSIS OF REINFORCED EMBANKMENT

M.HADA
Technical Research Institute,Fujita Co.
74 Ohdana-Cho Kohoku-ku
Yokohama 223 Japan

Abstract:The numerical analysis is carried out to many kinds of slope stabilities using the Rigid Body and Spring Model (hereafter called RBSM) proposed by Prof.Kawai(1977). This paper reports on the application of RBSM analysis to the retaining wall model test and the field embankment reinforced with polymer grids that RBSM is an effective analytical model for earth reinforcing mechanisms.

I .INTRODUCTION

In discrete limit analysis of reinforced soils, the interface between reifocing material and soil is modeled by the beam element and the plane element, each of them is defined by RBSM. Stress (τ , σ) of RBSM are transmitted by two springs (the shearing spring and the normal spring) distributed over the contact surface of two adjacent rigid elements. The discrete surface such as the interface can be easily assessed by properties of RBSM.

II .RETAINING WALL MODEL TEST

Fig.1 illustrates the retaining wall model test reinforced with vinyl sheets, in which the bottom of the wall at left side is hinged. Fig.2 shows the analytical result of slip lines inside the back-fill caused by moving of the wall. In the present analysis, it is assumed that yielding of the interface is defined by Mohr-Coulomb's criterion and plasitc strain is computed by the procedure using associated flow rule. Therefore, if the interface has become to be yielding,the strain of the adjacent sheet will not be able to increase. Test data and computed results of strain of vinyl sheets show that the occurrence of the slip in the upper sheet (the measured line number 21 corresponding to the computed result marked by ● in Fig.3).

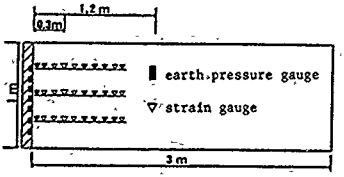


Fig.1 Sheets arrangement in the retaining wall model test (the wall is at left side)

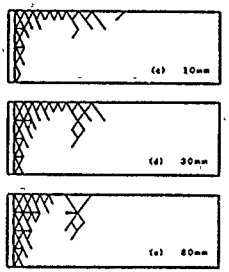


Fig.2 Slip line by the present analysis

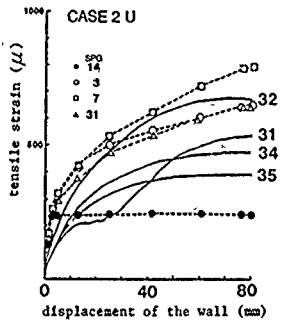


Fig.3 Occurrence of the slip in the upper sheet (Solid lines indicate test results)

III .REINFORCED EMBANKMENT WITH SURCHARGE

A 8.5m high and 70m long reinforced embankment with the 8m high surcharge fill was built on the stabilized foundation. The embankment reinforced with polymer grids has a slope of 1:0.3 and the upper fill has a slope of 1:0.8. Fill materials are mainly gravelly soils including the fine fraction. Therefore,it took for a long constructing period avoiding the settlement after completion.

Fig.4 shows the distribution of strain of polymer grids at three constructing stages. In this figure, strain marked by cicles are unreliable because one guage of pairs had become unstable. Fannin et al.(1988) presented that polymer grids laid in the field embankment were influenced by the long-term load. Kutara et al.(1988) showed that the long-term strain of polymer grids increased from about 0.5% to about 2% by their pull-out test. Therefore, it is considered that increasing of strain for the suspended term, without rainfall, caused by the long-term deformation of polymer grids themselves. So, the strain by the self-weight load at completion shoud be equal to the value given by deducting the long-term strain increment from the strain measured. Fig.5 shows that the analytical results agree well with the modified strain except the toe of slope.

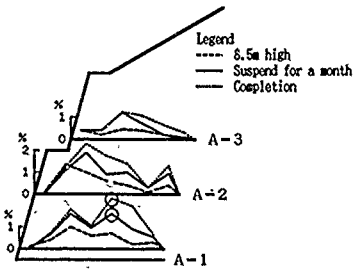


Fig.4 Distribution of strain of polymer grids

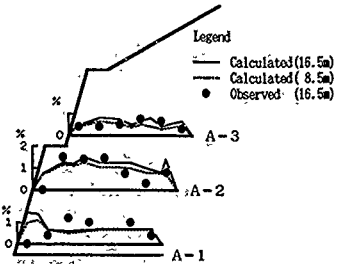


Fig.5 Comparison of the modified strain of polymer grids with computed results

IV .CONCLUSION

Computed results show explicitly the non-linearity of reinforced soils caused by the self-weight load and the reduction of earth pressure by reinforcing. However, deformation of reinforced embankments in field is more complicated. Especially, rainfall effect should be considered in Japan and other heavy rainfall countries. In this work, the rainfall effect was only calculated by increasing the unit weight of soils as well as current design methods.

ACKNOWLEDGEMENTS

The retaining wall model test data were quoted from the "Report on Experimental Studies of Earthquake Resisting Reinforcement of Abutment" for 1985 by Japan Highway Public Corporation. The authors are grateful for suggestion by persons concerned.

REFERENCES

Fannin,R.J. and Herman,S.(1988). Field Behaviour of Two Instrumented, Reinforced Soil Slopes:IGS on Theory and Practice of Earth Reinforcement, 277-282

Kawai,T.(1977).A New Discrete Model for Analysis of Solid Mechanics Problems: Seisan-Kenkyu, Vol. 29, No.4.

Kutara,K.,Aoyama,N.,Yasunaga,H.and Kato,K.(1988) Long- Term Pull-Out Tests of Polymergrids in Sand: IGS on Theory and Practice of Earth Reinforcement, 117-122

LIMIT ANALYSIS OF FRAMED STRUCTURES INCLUDING EFFECT OF FOUNDATION BY MEANS OF RIGID BODIES-SPRING MODEL

Masaaki MITO
PENTA-OCEAN Construction Co., LTD.
Tokyo 140, Japan

Nono TAKEUCHI
Department of Civil Engineering
Meisei University
Tokyo 191, Japan

Tadahiko KAWAI
Department of Electrical Engineering
Science University of Tokyo
Tokyo 162, Japan

Abstract- It has been already confirmed that the beam element of RBSM proposed by KAWAI is useful for limit analysis of framed structure. This paperpresents its application to the soil-structure interaction problem of framed structure. In this model effect of soil reaction distributed along individual beam component is integrated and it is replaced by three different springs lumped at the center of a given beam As a result, analysis of such non-linear problems can be simplified to great extent. The usefulness of this method will be discussed with some examples.

1.INTRODUCTION

Several mechanical models of foundation are proposed and they are expressed as combination of spring which obeys Hooke's law, dashpot which obeys Newton's law and slider which represents plasticity It is complicated to obtain stiffness matrix including the effect of visco-elastic, visco-plastic foundation explicitly by using the conventional beam elements. But a beam element of RBSM assumes to be rigid, so the effect of foundation pressure distributed along individual beam component is integrated and it is replaced by three different springs lumped at the center of a given beam. Considering this advantage, the characteristics of visco-elastic, visco-plastic foundation can be solved easily [1]

2. ANALYSIS OF BEAM ON A VISCO-ELASTIC FOUNDATION

2.1 Model of visco-elastic foundation

The visco-elastic medium is time dependent and its linear visco-elastic behaviour can be represented by combining the linearly elastic spring and the viscous dashpot filled with a liquid which obeys Newton's law of viscosity. In this paper a visco-elastic foundation is considered as shown in Fig.1. This is Standard solid model which Freudenthal and Lorsch used in an analysis of beam on a linear visco-elastic foundation.[2]

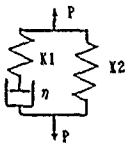


Fig.1 Standard solid model

2.2 Result of beam on visco-elastic foundation

Beam whoes section properties are uniform is analyzed in order to examine the usefulness of this model and a concentrated load is applied at the center of the beam and boundary condition is free at both edges. Section properties used are decided to become $(EI/K)/L^4 = 10^{-3}$ Where I is momentof inertia, E is modulus of elasticity and L is length of span. Time step interval used for calculation is $\Delta t/\tau = 0.5$ Where τ is η/k_1. With regard to the modeling, the length of half span isdivided into 5, 10, 15 and 20 elements in order to examine the relation between number of beam elements and convergency of solution. The deflection, bending moment and bearing pressure were compared with those of analytical solution carried by Sonoda et al. from $t/\tau = 0$ to $t/\tau = \infty$ as shown in Fig.2. In this case, number of elements were 10 Analytical solution is given by solid lines while result of the present analysis given by the mark ○. The deflection solved by this model is a slightly larger than analytical solution, but in general good agreement is observed. Both bending moment and bearing pressure are also in good agreement along the whole beam. The relation between the error of deflection at the midpoint and number of division at the optional time level from $t/\tau = 0$ to $t/\tau = \infty$ is shown in Fig. 3. The error of deflection decrease with increase mesh division, and same with the elapse of time [3]

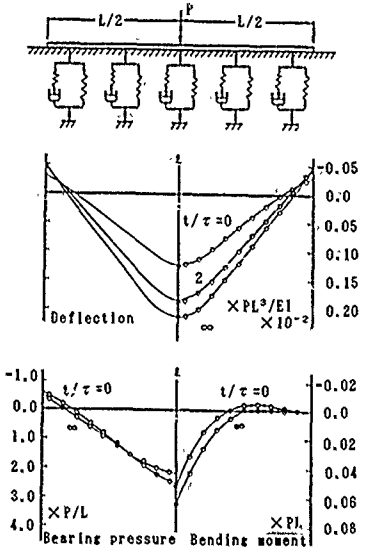


Fig.2 Beam on the visco-elastic foundation

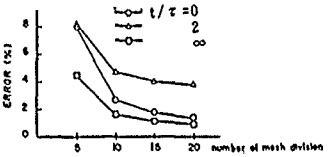


Fig 3 Convergence test of a beam element

3. LIMIT ANALYSIS OF FRAMED STRUCTURE

3.1 Model of visco-plastic foundation

The elasto-plastic medium shown in Fig 4 is not considered time dependent and it is assumed to become plastic instantaneously when it reaches yielding condition. The visco-plastic medium is considered time dependent but it ignores yielding of the foundation. The visco-plastic medium is considered as combination of elasto-plastic medium and visco-elastic one. In this paper the visco-plastic model of Bingham type shown in Fig.5 is used and this model express linear creep under yielding condition.

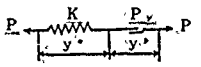


Fig.4 Elasto-plastic model

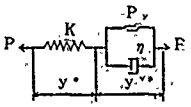


Fig 5 Visco-plastic model

3.2 Result of framed structure

Limit analysis model of framed structure made of uniform frame element is shown Fig 6. Also result of elasto-plastic analysis is shown in Fig.7. Number in the figure are order of collapse. At first the surface of the foundation is yield and plastic hinge appears at pile head. And then, yielding foundation advances downward. Finally plastic hinge appears under ground pile and the structure is collapsed. In this case, collapse load is 51 tf. Next, the results by means of visco-plastic method are described. Fig.8 shows the relation between displacement of pile head and time for three different loading velocity. Collapse occurred at 55 tf. This result is considered satisfactory because the calculated collapse load by the elasto-plastic analysis was 51 tf. For piles, the location and order of plastic hinge formation were the same as in case of the elasto-plastic analysis [4]

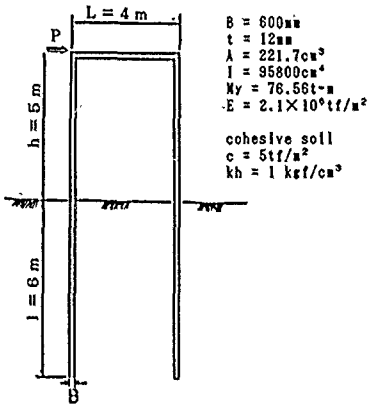


Fig 6 Analytical model

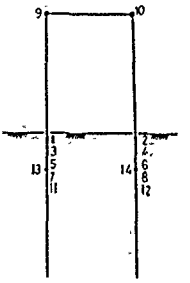


Fig.7 Collapse mode by using elasto-plastic model

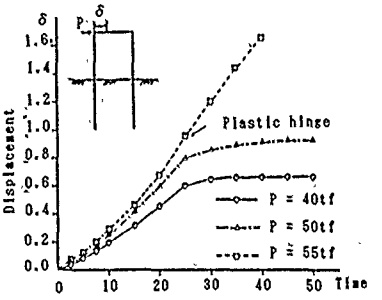


Fig.8 Time variation of displacement at the top of pile by using visco-plastic model

4. CONCLUSION

The framed structure including the effect of the foundation is analyzed by using discrete models proposed by KAWAI. The foundation is expressed as combination of springs, dashpot and sliders Analysis of the framed structure including the effect of visco-elastic, visco-plastic foundation has been considered intractable by using the existing finite element method. It is shown that the proposed method is very promising in analysis of such difficult soil-structure interaction problems.

REFERENCES

[1]T. Kawai New element models in discrete structural analysis, J. of the society of Naval Archiecture of Japan, Vol.141, 1977

[2]A. M. Freudenthal and H. G. Iorsch:The infinite elastic beam on a linear viscoplastic foundation, J. of the Engineering Mechanics Division, Proc. A S.C.E., Vol. 83

[3]N. Takeuchi, M. Mito and T. Kawai:An analysis of framed structure on a visco-elastic foundation by means new discrete models, J. of seisankenkyu, Institute of Industrial Science, Vol 34, No 11, 1982

[4]N. Takeuchi, M. Mito and T. Kawai An analysis of framed structure on a visco-plastic foundation by means new discrete models, J. of seisankenkyu, Institute of Industrial Science, Vol.34, No 11,

ON THE NEW DISCRETE ELEMENT MODEL OF VORONOI-POLYGON

TOSHIO YAMADA
Nonlinear Mechanics Inst.
Tokoh-White Buld. 2-11-9
Ebisu-Nishi, Shibuya
Tokyo 101 Japan

AND TADAHIKO KAWAI
Department of Engineering
Science University of Tokyo
1-3 Kagurazaka, Shinjyuku
Tokyo 162 Japan

Abstract-A new discrete element of Voronoi polygon is presented in this paper. This model is produced from Voronoi Regions using random points after studies of the crack pattern of the fractured or collapsed metallic materials, soil or rock mass. These problem can be solved only by using Kawai's RBSM, because the shape of the elements may be polygonal, the total number of sides may be 4-15.

From the results of analysis, it can be concluded that this new models give us the sets of suitable crack lines and their patterns whose charactaristics is of a Fractal Degree (D).[3]

I. INTRODUCTION

Frequently it is explained the collapse of the rock mass is discussed plotting the stress-strain curve. In situ, when we judge the safety on rock walls, for example, usually we might observe some cracking charactaristics of the slope, piller, roof or ground surface as the target. On that time, we can not get the stress-strain curve plot at the sites. It is believed that the strength can be predicted from the deformation pattern of the material mass. Insitu services, we evaluate the cause and effect about safety with ackrovledgments and knowhow from engineering standpoint of view.

II. COLLAPSE PATTERN OBSERVED IN THE FIELD STUDY

The linearments as the cracked lines are widely observed. Their length(L) related to frequency(N) as function

$$N(L) \propto L^{-D} \quad (D:\text{constant}) \qquad (1)$$

This function is adoptive on many facts from macro to micro and km to μm in length. This is a new experimental rule.

A. Geological linearments of Satellite-Photos: When these lines control frequency relationship in length, we can get the very simple rule. It's in fractals. The line length is about 500-Nx1000m (N>1).

B. Geophisical hypothesis tell us another information that is conserning a limitted mass block. One is the wide dimension in about 100km flat, another is about the limitted depth in 30-50km, called Mohorovicic discontinuity surface.

C. After crushing, the distribution size of the rock about the accumulate weight ratio over sieve has a linearity in RRB(Rosin-Rammler-Benett)-Plot.[1,2] The relation is given as follow.

$$\log\log(100/Wt) \propto \log(d) \qquad (2)$$

in size(d) and accumulated weight percents(Wt) over sieve.

D. Crack lines in rock wall are appearently arbitrary. But their frequency of length are in logalismic normal distribution. Sum of frequency on length present fractal relation about the density of distribution on square net.

E. Some micro-cracks are remained in minerals of rocks including some gas or liquid which were created in geological old period. They are the open cracks or adhesive cracks and their size are 20μ and 15μm in maximum. Frequency of traced length of them have fractal relationship in length.

III. VORONOI REGION

A. Definition

We can get a set of the triangle net in a area connecting arbitrary points on a flat plane. A edge of Voronoi polygon is perpendicular at the center of the edge of a triangle to it.(Fig.1) A Voronoi point is the cross point of this perpendicular lines. An area is devided into Voronoi regions by this lines. The Voronoi points define the circumcenters of Delaunay's triangle in coordinates. The count of initial points is same as the count of new polygons to produce. The Voronoi-regions are polygonal. Now if we have arbitrary points, it is easy to convert the net of polygons on them exactly. Also arbitrary points can be obtained from the experimental datas.

B. Charactaristics

Voronoi polygons have the many unique charactaristics.

1) Count of tops on a polygon is six in anticipation.
2) This polygon is convex shape completely.
3) Directions of boundary edges of polygons which are generated crack points by random numbers in uniform, are in uniform.
4) Frequency of length of edges in polygons have a Fractal Coefficient.(Fig.3)
5) Frequency of size in polygons has the fractals too.

Charactaricities 4),5) are very interesting and important.

IV. MAKING THE MODEL

How to make the polygonal net on our limited plane ? After considering about experimental rules, we had developed several techniques include geometrical autogenerating tools. Usually it's generated voronoi-polygons by program "POLYG". If necessary, it can get to overlay some stright lines as the pre-cut slits or special cracks on the model.

V. SOLVER

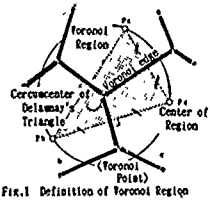


Fig.1 Definition of Voronoi Region

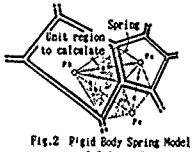


Fig.2 Rigid Body Spring Model with Polygonal shape

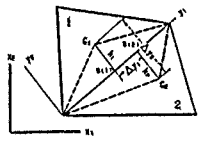


Fig.4 To Improve Accuracy

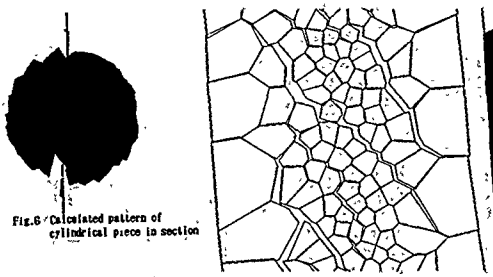
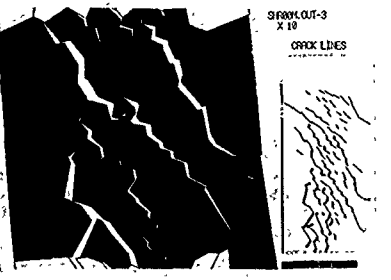


Fig.6 Calculated pattern of cylindrical piece in section

a. Calculated elements(a part)



b. Crack pattern in arbitrary scale

Fig.5 Calculated patterns

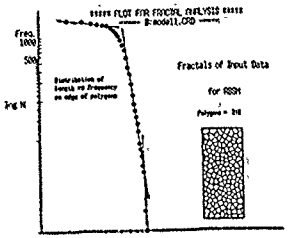


Fig.3 Fractals in Polygonal Net

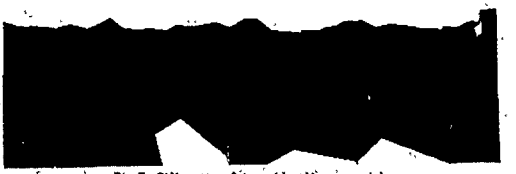


Fig.7 Silhouette of top ridge line on model in exaggerate scale (likes as landscope)

A. Consideration. The geometrical dual relation between Delaunay's triangle and Voronoi region is more suitable to KAWAI's RBSM.(Fig.2) On the Voronoi polygons, a boundary edge cross in perpendicular to a edge of Delaunay's triangle. On RBSM, if the optional point to involve force set the center of polygon, we can get the most agreeable condition to analize. This technique can not be used in case of FEM.

B. Improve accuracy. Referring to Fig.4.

$$\frac{\partial u}{\partial x_j} = \frac{\partial u}{\partial y_1}\frac{\partial y_1}{\partial x_j} = \frac{\partial u}{\partial y_1}\frac{\partial y_1}{\partial x_j} + \frac{\partial u}{\partial y_2}\frac{\partial y_2}{\partial x_j} \qquad (3)$$

When $(\Delta y_1, \Delta y_2)$ are the distances between (G_1, G_2) and edge line,

$$\frac{\partial u}{\partial x_j} = \frac{u_{(1)} - u_{(2)}}{\Delta y_1} + \frac{\partial y_1}{\partial x_j} + \frac{u_{(1)} - u_{(2)}}{\Delta y_2}\frac{\partial y_2}{\partial x_j} \qquad (4)$$

$$\Delta y_2 \gg \Delta y_1 \geqq 0$$

When evaluting position of strain energy balance on the boundary edge of voronoi regions set the summit of Delaunay's triangle, perfectly $\Delta y_1 = 0$. Then,

$$\frac{\partial u}{\partial x_j} = \frac{u_{(1)} - u_{(2)}}{\Delta y_2}\frac{\partial y_2}{\partial x_j} \qquad (5)$$

So the accuracy improve unconditionally in spite of geometrical arrange of the constructed net. It is some limited cases of condition that is "$\Delta y_1 = 0$", on which elements are in regular triangle, right-angle triangle. The rectanglar must be separate as like Union-Jack flag, square and regular hexagon etc. These limited net models are very difficult to generate exactly. Voronoi polygonal net are very easy and exact to construct and the best model to get the best accuracy completely.

VI. ANALYSIS

A. Create crack branch and their fractals(Fig.5)

We always conduct material testing to survey the strength and charactaristics of materials used. On such cases, we observe crack lines initiated on the pieces as like branches of tree.

After shearinng calculation, it happend u mount of plastic hinge and broken boundary lines which have a fractal degree with their length and frequency. The pattern of these lines show likes tree shape, which has a few trunk, bough, twig and leaves much. Simulated material is concrete mortal. Finite elements to calculate are natural random arrayed polygons. Just force create his crack ways with hisself.

B. Collapse of cylindrical test piece(Fig.6)

We frequently studied shearing strength. This simulation was made under plain strain condition at a section. The loads set on two pin points by the top and bottom. Loading balance is not so exactly, so the deformation of elements in model are unbalance. After the result, broken elements set their position in natural on arbitorary scaling.

C. Collapse model of Earth Crust(Fig.7)

Under the horizontal pressure, a rock mass model of rectangular body deform with hinge and crack lines. Upper edge line shows a modal pattern. This deformation images a landscope in perspectives. It looks far mountain ridge.

VII. CONCLUSION

This paper presents a new finite element model of polygonal shape. They can solve only on RBSM. Generally, it can't solve this polygonal net problems by FEM.

It is described that the relationship of dual relation between Delaunay and Voronoi Regions is suitable for KAWAI's RBSM. The length of each edges beyond Voronoi Polygons have the fractal degree. So, after loaded analysis, collapsed hinge lines and broken mass show fractals about charactaristics and shape of model on outlines.

Simulated crack pattern is presented which is analogous to the patterns of linearment as like as photo-image.

> References <

1) Rosin,P.,E.,Rammler und K.Spering:Reichskohlenrat, "Bericht C.52", Berlin, 1933
2) Benett,J.,G.: J.Inst.Fuel,10, pp22-33, 1936
3) Mandelbrot,B.,B.,:The Fractal Geometry of Nature, Freemann, San Fransisco, 1982

A NONCONFORMING FINITE ELEMENT METHOD FOR THE REISSNER-MINDLIN PLATE BENDING MODEL

Rolf Stenberg, Teemu Vihinen
Faculty of Mechanical Engineering
Helsinki University of Technology
02150 Espoo, Finland

Leopoldo P. Franca
Laboratório Nacional de Computação Científica
Rua Lauro Müller 455
22290 Rio de Janeiro, RJ - Brazil

Abstract - We present a new linear nonconforming finite element method for Reissner-Mindlin plates. For the element we have proved optimal order error estimates which are uniformly valid with respect to the plate thickness. This means that the element will never "lock". We give numerical examples which show that this is the case.

I. THE ELEMENT

The weak form of the Reissner-Mindlin plate model is the following (cf. [4]): Find the deflection $w \in H_0^1(\Omega)$ and the rotation $\beta \in [H_0^1(\Omega)]^2$ such that

$$E(\beta, w) \le E(\psi, v) \quad \forall (\psi, v) \in [H_0^1(\Omega)]^2 \times H_0^1(\Omega),$$

with the energy defined as

$$E(\psi, v) = \frac{t^3}{2}\, a(\psi, \psi) + \frac{G\kappa t}{2} \int_\Omega |\psi - \nabla v|^2 \, dx - (f, v),$$

where G is the shear modulus and κ is the shear correction factor. The thickness of the plate is denoted by t. The bilinear form a is given by

$$a(\beta, \psi) = \frac{E}{12(1-\nu^2)} \int_\Omega [(1-\nu)\, \epsilon(\beta) : \epsilon(\psi) + \nu \operatorname{div} \beta \operatorname{div} \psi] \, dx,$$

with E and ν denoting the Young modulus and Poisson ratio, respectively. For simplicity we present the case of a clamped plate.

The locking of finite element methods occurs for "thin" plates and hence it is customary to consider a sequence of problems obtained when assuming that the transversal load is of the form

$$f = t^3 g \tag{1}$$

with g fixed. This ensures that the plate model has a finite limit solution when $t \to 0$. We will make this assumption below.

In this note we will present a new finite element method introduced in [3] and independently in [2] by Durán, Ghioldi and Wolanski. The method can be viewed as a modification of a recent method by Arnold and Falk [1]. In the method the deflection is approximated with linear nonconforming elements: we let $\mathcal{C}_h$ be a triangulation of the domain and define

$$W_h = \{\, v \in L^2(\Omega) \mid v|_K \in P_1(K),\ K \in \mathcal{C}_h \text{ and } v \text{ is continuous at midpoints of element edges and vanishes at midpoints of boundary edges} \,\}.$$

For the rotation we use the standard space of continuous piecewise linear functions

$$B_h = \{\, \beta \in [H_0^1(\Omega)]^2 \mid \beta|_K \in [P_1(K)]^2,\ K \in \mathcal{C}_h \,\}.$$

The discretization is then defined as: find $w_h \in W_h$ and $\beta_h \in B_h$ such that

$$E_h(\beta_h, w_h) \le E_h(\psi, v) \quad \forall (\psi; v) \in B_h \times W_h,$$

with the following modified energy expression

$$E_h(\psi, v) = \frac{t^3}{2}\, a(\psi, \psi) + \sum_{K \in \mathcal{C}_h} \frac{G\kappa t^3}{2(t^2 + \alpha h_K^2)} \int_K |\mathrm{R}_h(\psi - \nabla v)|^2 \, dx - (f, v).$$

Here h_K denotes the diameter of the element $K \in \mathcal{C}_h$. The operator R_h is defined as the L^2-projection onto the piecewise constant vectors, i.e. we have

$$(\mathrm{R}_h \psi)|_K = \frac{1}{\operatorname{area}(K)} \int_K \psi \, dx, \qquad K \in \mathcal{C}_h.$$

The parameter α is positive and fixed. In practice the inclusion of the reduction operator simply means that the local stiffness matrix is calculated with the one point integration formula. For our method we have proved the following error estimate [3].

THEOREM. *Suppose that Ω is convex and that the load f is in $L^2(\Omega)$ and satisfies (1). Then there is a positive constant C, independent of the plate thickness t, such that*

$$\|w - w_h\|_0 + \|\beta - \beta_h\|_0 \le C h^2 \|g\|_0.$$

For the details of the error analysis and for a discussion of the advantage of the method compared with the method of Arnold and Falk we refer to [3].

II. NUMERICAL EXAMPLES

We will give results of computations for a clamped and simply supported square plate. We consider both a uniform and concentrated center point load. The symmetry of the solution is utilized and only one quarter of the plate is discretized. The type of meshes used is seen from Figure 1.

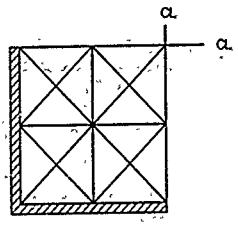


FIGURE 1. The finite element mesh with $h = 1/2$.

The side length of the quarter plate is chosen equal to unity and we let $\nu = 0.3$. We have chosen $t = 0.01$ and compare our results with the exact classical Kirchhoff solution. We use the "hard" simply supported boundary conditions. In principle the parameter α can be chosen arbitrary as long as it is positive, but is clear that the solution will depend on the choice. The best results we have obtained for α in the range $0.05 - 0.3$. In the calculations presented below we have chosen $\alpha = 0.1$. We denote by M_x the exact normal moment and by M_x^h the approximation calculated from the discrete rotation β_h.

Our numerical results clearly show that the method performs very well. We have also indicated the convergence rates, i.e. the powers of the mesh length h by which the L^2-errors decrease, and we see that the calculations confirm the error estimates proven mathematically.

Some further numerical results and comparisons with some other linear elements are presented in [5].

TABLE 1. The normalized center deflection w_h(center)/w(center) for the uniform load.

1/h	Clamped	Simply supported
2	1.0570	1.0268
4	1.0121	1.0077
8	1.0043	1.0023

TABLE 2. The error $\|w - w_h\|_0/\|w\|_0$ for the uniform load.

1/h	Clamped	Simply supported
2	0.0663	0.0477
4	0.0205	0.0117
8	0.0044	0.0026
Conv. rate	1.96	2.10

TABLE 3. The error $\|M_x - M_x^h\|_0/\|M_x\|_0$ for the uniform load.

1/h	Clamped	Simply supported
2	0.4379	0.2002
4	0.2285	0.1001
8	0.1149	0.0501
Conv. rate	0.97	1.00

TABLE 4. The normalized center deflection w_h(center)/w(center) for the concentrated center load.

1/h	Clamped	Simply supported
2	1.0755	1.0744
4	1.0418	1.0305
8	1.0193	1.0119

TABLE 5. The error $\|w - w_h\|_0/\|w\|_0$ for the concentrated center load.

1/h	Clamped	Simply supported
2	0.0710	0.0403
4	0.0188	0.0114
8	0.0057	0.0033
Conv. rate	1.82	1.81

TABLE 6. The error $\|M_x - M_x^h\|_0/\|M_x\|_0$ for the concentrated center load.

1/h	Clamped	Simply supported
2	0.5851	0.4419
4	0.2993	0.2161
8	0.1424	0.1015
Conv. rate	1.02	1.06

REFERENCES

1. D.N. Arnold and R.S. Falk, A uniformly accurate finite element method for the Reissner-Mindlin plate. SIAM J. Num. Anal. 26 (1989) 1276-1290
2. R. Durán, A. Ghioldi and N. Wolanski, A finite element method for the Mindlin-Reissner model. SIAM J. Num. Anal. To appear
3. L.P. Franca and R. Stenberg, A modification of a low-order Reissner-Mindlin plate bending element. J.R. Whiteman (Ed.), The Mathematics of Finite Elements and Applications VII. MAFELAP 1990. Academic Press. To appear
4. T.J.R. Hughes, The Finite Element Method. Linear Static and Dynamic Analysis. Prentice-Hall 1987
5. R. Stenberg and T. Vihinen, Calculations with some linear elements for Reissner-Mindlin plates. Proceedings of the European Conference on New Advances in Computational Structural Mechanics. April 2-5, 1991, Giens, France. To appear

On locking effects in the finite element method

Manil Suri
Department of Mathematics and Statistics
University of Maryland Baltimore County
Baltimore, MD 21228, USA

1. Introduction

The formulation of a number of partial differential equations involves a parameter-dependency which arises out of physical considerations. For example, plate and shell models involve 't', the thickness, heat transfer through anisotropic materials involves 'k', the ratio of conductivities in different directions, and elasticity problems involve 'ν', the Poisson ratio. *Locking* is the term given to the deterioration in numerical schemes which may occur when the parameter is close to a limiting value. ($t \to 0, k \to 0, \nu \to 0.5$). For instance, it is well-known that the use of the h-version FEM with piecewise linear elements for nearly incompressible materials (i.e. for ν close to 0.5) is ineffective, precisely due to locking (see [1] for other examples of locking).

A *robust* method is one which guarantees a certain rate of convergence uniformly with respect to a parameter. Various robust schemes have been suggested for different problems involving locking - for example, mixed methods, higher order h-version methods and the p-version. In [2], we have developed a general mathematical theory for locking and robustness and their quantitative assessment, which can be helpful in the analysis of such methods. In [3], this theory has been applied to a specific example, that of Poisson ratio locking, and the locking and robustness of various schemes has been analysed.

Our goal in this paper is to present some of the theory from [2], which we do in terms of the specific problem of Poisson ratio locking. It should be kept in mind that most of the definitions and theorems we state in Section 2 are applicable to general parameter-dependent problems as well (see [2]). In Section 3, we present some theorems from [3] for the robustness of various schemes when the Poisson ratio is close to 0.5.

2. Locking and robustness

We consider the elasticity equations

$$-\Delta \vec{u}_\nu - (1-2\nu)^{-1}\operatorname{grad}\operatorname{div}\vec{u}_\nu = 0 \quad \text{in } \Omega \tag{2.1}$$

$$\sum_{j=1}^{2}(\epsilon_{ij}(\vec{u}_\nu) + \delta_{ij}\nu(1-2\nu)^{-1}\operatorname{div}\vec{u}_\nu)n_j = g_i \quad \text{on } \Gamma, \tag{2.2}$$

where $i = 1,2$, ϵ is the usual strain tensor, and $\vec{g}$ satisfies a compatibility condition. Here, ν, the Poisson ratio, lies in $S = [0, 0.5)$. Let $V = \left(H^1(\Omega)\right)^2$ and

$$B_\nu(\vec{u},\vec{v}) = \int\int_\Omega \left(\sum_{i,j=1}^{2}\epsilon_{ij}(\vec{u})\epsilon_{ij}(\vec{v}) + \nu(1-2\nu)^{-1}\operatorname{div}\vec{u}\operatorname{div}\vec{v}\right)dx$$

with the energy norm $|||u|||_{E,\nu} = \left(B_\nu(u,u)\right)^{1/2}$.

In the limiting case, as $\nu \to 0.5$, equations (2.1)–(2.2) lead to a Stokes' problem and the solution satisfies the incompressibility constraint

$$D\vec{u} = \operatorname{div}\vec{u} = 0. \tag{2.3}$$

In the sequel, we take $H = \left(H^{k+1}(\Omega)\right)^2$ and denote $|||\cdot|||_\nu$ by

$$|||\vec{u}|||_\nu^2 = |||\vec{u}|||_{E,\nu}^2 + |\vec{u}|_{k+1,\Omega}^2 \tag{2.4}$$

Also, define $H^B = \{u \in H \mid |||u|||_B \le K\}$ and $H_\nu^B = \{u \in H \mid |||u|||_\nu \le K\}$. (Here, K may represent different values but it will always be bounded by K_0, a constant independent of ν.)

For each $N \in \mathcal{N}$, N unbounded, let $V^N \subset V$ be a finite element subspace of dimension N. Given any "exact solution" $u_\nu \in H_\nu$, we may now define the "finite element approximation" $u_\nu^N \in V^N$ by

$$B_\nu(u_\nu^N, v) = B_\nu(u_\nu, v) \qquad \forall v \in V^N \tag{2.5}$$

The set $\mathcal{F} = \{V^N\}$ therefore describes an *extension algorithm*, i.e. a rule for increasing N (and hence the accuracy). Let E_ν be an error functional, $E_\nu(v) = |||v|||_\nu$ or $|||v|||_{E,\nu}$. We assume that

$$C_1F_0(N) \le \sup_{u\in H^B}\ \inf_{v\in V^N} |||u-v|||_\nu \le C_2F_0(N)$$

and for $\nu \in [0,b]$, $b < 0.5$,

$$C_1F_0(N) \le \sup_{u_\nu \in H_\nu^B} E_\nu(u_\nu - u_\nu^N) \le C_2F_0(N) \tag{2.6}$$

where C_1, C_2 are positive constants independent of N, ν. Here, $F_0(N) \to 0$ as $N \to \infty$ and is independent of ν. Then we have the following definitions.

Definition 2.1 Let $\lim_{N\to\infty} f(N) = \infty$. The extension algorithm $\mathcal{F}$ shows locking of order $f(N)$ for the family of problems (2.5), $\nu \in [0, 0.5)$, with respect to the solution sets $\{H_\nu\}$ and error measures $\{E_\nu\}$ iff

$$0 < \lim_{N\to\infty} \sup\left(\sup_{\nu\le S}\left(\sup_{u_\nu\in H_\nu^B} E_\nu(u_\nu - u_\nu^N)\right)\left(F_0(N)f(N)\right)^{-1}\right) = M < \infty \tag{2.7}$$

For the case that M is bounded (respectively, infinite), we say that the order of locking is at most (respectively, at least) $f(N)$. If (2.7) holds with $f(N) \equiv 1$, then we say that $\mathcal{F}$ is free from locking.

Definition 2.2 Let $g(N) \to 0$ as $N \to \infty$. The extension algorithm $\mathcal{F}$ is robust with uniform order $g(N)$ for the family of problems (2.5), $\nu \in [0, 0.5)$, with respect to the solution sets $\{H_\nu\}$ and error measures $\{E_\nu\}$ iff

$$\lim_{N\to\infty} \sup\left(\sup_{\nu\le S}\left(\sup_{u_\nu\in H_\nu^B} E_\nu\left(u_\nu - u_\nu^N\right)\right)\left(g(N)\right)^{-1}\right) = M < \infty$$

We see from the above definitions that if $f(N)$ is such that

$$f(N)F_0(N) = g(N) \to 0 \quad \text{as } N \to \infty$$

then $\mathcal{F}$ shows locking of order $f(N)$ iff it is robust with maximum uniform order $g(N)$.

Let H_β denote the limit of the sets H_ν, consisting of elements $u \in H$ satisfying (2.3). Define $H^\beta_\beta = E_\beta \cap H^\beta$. Then the following condition has been proven for the problem (2.1),(2.2) in [3].

Condition (a) For any $u_\nu \in H^\beta_\nu$, there is a $u_\beta \in E^\beta_\beta$ (u_β depending on u_ν) such that

$$\|u_\nu - u_\beta\|_E \le C(1-2\nu)^{1/2} \tag{2.8}$$

with C independent of ν and u_ν.

The following theorem shows the usefulness of the set H^β_β.

Theorem 2.1 *Let $E_\nu(u) = \|u\|^2_{E,\nu}$ or $\|u\|^2_{E,\nu}$. Let Condition (a) hold. Then $\mathcal{F}$ shows locking of order $f(h)$ with respect to E_ν iff*

$$g(N) = \sup_{u \in E^\beta_\beta} \inf_{v \in V^N, D v = 0} \|u - v\|^2_{E,\nu} \approx C F_\beta(N) f(N).$$

and is free from locking iff the above holds with $f(N) = 1$. Moreover, $\mathcal{F}$ shows locking of order $f(N)$ in the V norm iff it shows locking of order $f(N)$ in the energy norm.

The advantage of Theorem 2.1 is that one reduces the question of locking to a question of approximability alone, involving $g(N)$. In fact, for our problem, the following equivalent characterization of $g(N)$ is very useful.

Theorem 2.2 *Let $g(N)$ be as defined in Theorem 2.1. Then $g(N)$ is equivalent to*

$$\tilde{g}(N) = \sup_{\phi \in \tilde{H}^\beta} \inf_{\chi \in \tilde{W}^N} \|\phi - \chi\|^2_{2,\Omega}$$

where $\tilde{H}^\beta = \{\phi \in H^{1+2}(\Omega) \mid \|\phi\|^2_{1+2,\Omega} \le K\}$ and $W^N = \{\chi \mid \nabla \times \chi \in V^N\}$.

The above theorem is used, for instance, in the proof of Theorem 3.1 in the next section (see[3]).

3. Robustness results

We now consider problem (2.1),(2.2) over $\Omega = (-1,1)^2$. Let $\{T^h_1\}$ and $\{T^h_2\}$ be the uniform rectangular and uniform triangular meshes shown in Figure 3.1(a) and (b) respectively. For any set $S \subset \mathbb{R}^2$, denote by $\mathcal{P}^1_p(S)\big(\mathcal{P}^2_p(S)\big)$ the set of polynomials of total degree (degree in each variable) $\le p$ and let $\mathcal{P}^3_p(S) = \mathcal{P}^1_p(S) \oplus \{x^p y, x y^p\}$. Define

$$V^i_{p,h} = \{u \in C^0(\Omega), \quad u|_S \in \mathcal{P}^i_p(S) \quad \forall S \subset T^h_i\}$$

for $i = 1,2,3$ where $T^h_3 = T^h_2$. Then we have ([3],[4])

Theorem 3.1 *Let $\mathcal{F}$ be an h-version extension algorithm for the problem (2.5) using subspaces $V^N = (V^i_{p,h})^2$ with p fixed. Then with $H, \|\cdot\|, \|\cdot\|_\nu$ as above (k large enough), the following is true for locking in the energy norm*

$V^1_{p,h}$ For $p \le 3$, the order of locking is $O(h^{-1})$ and the order of robustness is $O(h^{p-1})$. For $p \ge 4$, there is no locking and the order of robustness is $O(h^p)$.

$V^2_{p,h}$ For $p \ge 1$, the order of locking is $O(h^{-1})$ and the order of robustness is $O(h^{p-1})$

$V^3_{p,h}$ For $p \le 2$, the order of locking is $O(h^{-1})$ and the order of robustness is $O(h^{p-1})$. For $p \ge 3$, the order of locking is $O(h^{-2})$ and the order of robustness is $O(h^{p-2})$.

The above shows that locking cannot be avoided for $p \le 3$ for the h-version. Also, using rectangular elements leads to locking for all p.

In contrast, it was shown in [5] that the p-version (using straight-sided triangles) leads to uniform robustness of order that is optimal up to an arbitrary $\epsilon > 0$. In [3], we show that using our formulation, it is possible to prove this optimal order without the ϵ, both for triangles and rectangles, so that locking is completely eliminated. For the p-version, however, the use of curvilinear elements is indispensable in most cases. In this connection, we have the following theorem.

Theorem 3.2 *Let $\mathcal{F}$ be the p-version for the problem (2.5) using curvilinear elements S_i (triangles and quadrilaterals) which can be mapped onto the standard triangle or square using rational mappings B_i. Then $\mathcal{F}$ is free from locking and is uniformly robust with order $(p-s)^{-(k-1)}$ where $s \ge 0$ depends upon the mappings B_i but is independent of p.*

When the mappings are affine, we may take $s = 0$. For additional related results, see [3]

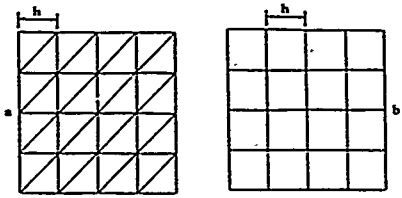


Figure 3.1 Uniform meshes (a)Rectangular (b)Triangular

Acknowledgement This work was partially supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant Number AFOSR 89-0252.

References

[1] K. J. Bathe and D. R. J. Owen (eds.), *Reliability of Methods for Engineering Analysis*, Pineridge Press, Swansea, U.K. (1986).

[2] I. Babuška and M. Suri, On locking and robustness in the finite element method, Tech. Note BN 1112 (1990) Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA.

[3] I. Babuška and M. Suri, Locking effects in the finite element approximation of elasticity problems, Tech. Note BN 1119 (1990) Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA.

[4] L. R. Scott and M. Vogelius, Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials, RAIRO Math. Mod. and Num. Anal. 19 (1985) 111-143.

[5] M. Vogelius, An analysis of the p-version of the finite element method for nearly incompressible materials, Numer. Math. 41 (1983) 39-53.

Asymptotic, Homogenization and Galerkin methods in Three-Dimensional Beam theory

M.L. Mascarenhas
Dep. Matemática, FCL and CMAF
Av. Prof. Gama Pinto 2,
1699 Lisboa Codex, Portugal.

L. Trabucho
CMAF
Av. Prof. Gama Pinto 2,
1699 Lisboa Codex, Portugal.

Abstract. - In this work we consider a three-dimensional linearized elasticity beam model and study different approximations of the three-dimensional solution using asymptotic, homogenization and Galerkin methods. We show that the choice of the basis functions of the Galerkin approximation, commutes with the homogenization process but the same does not hold for the final approximation of the three-dimensional solution.

I. INTRODUCTION

The summation convention on repeated indices will be used. Latin indices (i,j,k,...) shall take values in {1,2,3} and Greek indices ($\alpha,\beta,\gamma,\ldots$) shall take values in {1,2}.

Let (e_i) be the canonical basis of $\mathbb{R}^3$ and let ω be a domain in the plane spanned by the vectors e_α. Let $|\omega| = 1$ and let $\gamma = \partial\omega$ denote its boundary. Given $L \in \mathbb{R}$, $L > 0$, define :

$$\Omega = \omega \times]0,L[, \quad \Gamma_0 = \omega \times \{0,L\}, \quad \Gamma_1 = \gamma \times]0,L[.$$

Consider $Y = [0,1] \times [0,1]$ and let T be an open subset of Y, with a regular boundary ∂T, such that $\bar{T} \subset$ *int* Y. Define $Y^* = Y - T$ and let χ be its characteristic function, periodically extended to all the plane generated by (e_α). Let Θ be the area of Y^*, that is, $\Theta = |Y^*|$. Given $\delta \in \mathbb{R}, \delta > 0$, consider χ^δ given by $\chi^\delta(x) = \chi(\frac{x}{\delta})$, for all $x \in \mathbb{R}^2$. χ^δ defines a subset of $\mathbb{R}^2$ with a periodic structure having δY^* as basic cell. Suppose δ is a small parameter. If χ_ω represents the characteristic function of ω, $\chi^{\delta,1} = \chi_\omega \chi^\delta$ defines a subset $\omega^{\delta,1}$, of ω, obtained by perforating ω with holes δT, with a periodicity δY. To avoid a nonregular boundary, only holes whose closure does not touch γ are to be considered, *i.e.* , whenever the boundary of a hole intersects γ we will consider $\chi^{\delta,1} = 1$ in all the respective cell. This adjusted function $\chi^{\delta,1}$ defines a set that will be represented by $\omega^{\delta,1}$, with boundary $\gamma^{\delta,1}$. Define

$$\Omega^{\delta,1} = \omega^{\delta,1} \times]0,L[, \quad \Gamma_0^{\delta,1} = \omega^{\delta,1} \times \{0,L\}, \quad \Gamma_1^{\delta,1} = \gamma^{\delta,1} \times]0,L[.$$

Given $\varepsilon \in \mathbb{R}, \varepsilon > 0$, a dimensionless parameter that may be as small as we please, consider the homothety $j^\varepsilon(x_1,x_2) = (\varepsilon x_1, \varepsilon x_2) = (x_1^\varepsilon, x_2^\varepsilon)$, for all $(x_1,x_2) \in \mathbb{R}^2$ and denote

$$\omega^\varepsilon = j^\varepsilon(\omega), \quad \gamma^\varepsilon = j^\varepsilon(\gamma), \quad \omega^{\delta,\varepsilon} = j^\varepsilon(\omega^{\delta,1})$$
$$\partial T^{\delta,\varepsilon} = j^\varepsilon(\partial T^{\delta,1}), \quad \gamma^{\delta,\varepsilon} = j^\varepsilon(\gamma^{\delta,1}),$$
$$\Omega^\varepsilon = \omega^\varepsilon \times]0,L[, \quad \Gamma_0^\varepsilon = \omega^\varepsilon \times \{0,L\}, \quad \Gamma_1^\varepsilon = \gamma^\varepsilon \times]0,L[,$$
$$\Omega^{\delta,\varepsilon} = \omega^{\delta,\varepsilon} \times]0,L[, \quad \Gamma_0^{\delta,\varepsilon} = \omega^{\delta,\varepsilon} \times \{0,L\}, \quad \Gamma_1^{\delta,\varepsilon} = \gamma^{\delta,\varepsilon} \times]0,L[.$$

Let $n = (n_i)$ (respectively $n^\varepsilon = (n_i^\varepsilon)$) denote the outward unit normal vector to $\partial\Omega$ or to $\partial\Omega^{\delta,1}$ (respectively $\partial\Omega^\varepsilon$ or $\partial\Omega^{\delta,\varepsilon}$) and define the following differential operators, depending on ε, $\partial_\alpha^\varepsilon = \frac{\partial}{\partial x_\alpha^\varepsilon}, \partial_n^\varepsilon = \frac{\partial}{\partial n^\varepsilon}, \Delta^\varepsilon = \partial_{\alpha\alpha}^\varepsilon, div^\varepsilon(v) = \frac{\partial v_\alpha}{\partial x_\alpha^\varepsilon}$.

II. THE PROBLEM UNDER STUDY

We consider a beam occupying volume $\Omega^{\delta,\varepsilon}$ made of an homogeneous isotropic linearly elastic material with Lamé's constants λ and μ, both independent of δ and ε. Let $f_i^\varepsilon \in L^2(\Omega^{\delta,\varepsilon})$ and $g_i^\varepsilon \in L^2(\Gamma_1^{\delta,\varepsilon})$ be the i^{th} components of the volume density of applied body forces and surface density of applied surface tractions, respectively. With no loss of generality we consider that the system of applied forces does not depend on the parameter δ. We also admit that $g_i^\varepsilon(x^\varepsilon) = 0$ if $x^\varepsilon \in \partial T^{\delta,\varepsilon} \times]0,L[$. Consider the space of admissible displacements

$$V(\Omega^{\delta,\varepsilon}) = \{v = (v_i) \in [H^1(\Omega^{\delta,\varepsilon})]^3 : v = 0 \text{ on } \Gamma_0^{\delta,\varepsilon}\},$$

equipped with the usual H^1-norm and the following three-dimensional elasticity problem in $\Omega^{\delta,\varepsilon}$;

$$\begin{cases} u^{\delta,\varepsilon} \in V(\Omega^{\delta,\varepsilon}), \\ \int_{\Omega^{\delta,\varepsilon}} [2\mu e^\varepsilon(u^{\delta,\varepsilon}) + \lambda tr e^\varepsilon(u^{\delta,\varepsilon}) Id] e^\varepsilon(v^{\delta,\varepsilon}) d\Omega^{\delta,\varepsilon} = \\ = \int_{\Omega^{\delta,\varepsilon}} f^\varepsilon \cdot v d\Omega^{\delta,\varepsilon} + \int_{\Gamma_1^{\delta,\varepsilon}} g^\varepsilon \cdot v d\Gamma_1^{\delta,\varepsilon}, \quad \forall v \in V(\Omega^{\delta,\varepsilon}), \end{cases}$$

where, for each $v \in V(\Omega^{\delta,\varepsilon})$, $e^\varepsilon(v)$ represents the linearized elasticity strain tensor and Id the second order identity tensor. It is a classical result that this problem has a unique solution.

We are interested in the behaviour of $u^{\delta,\varepsilon}$, when δ and ε tend to zero.

III. THE GALERKIN APPROACH

An equivalent formulation of the elasticity problem posed now in a fixed (with respect to ε) domain $\Omega^{\delta,1}$ is considered. This is achieved using the same type of transformation as in the asymptotic expansion method for beams. Then, the transformed displacement field $u^\delta(\varepsilon)$ is approximated by $\pi_N^\delta(\varepsilon)$ the projection of $u^\delta(\varepsilon)$ onto the approximation space $V_N(\Omega^{\delta,1})$ of $V(\Omega^{\delta,1})$. The following result holds .

Theorem 1. *Let $f_i(x) = 0$ if $x \in \Omega^{\delta,1}$; $g_\alpha(x) = 0$ if $x \in \Gamma^{\delta,1}$ and $g_3(x) = 0$ if $x \in \partial T^{\delta,1} \times]0,L[$. Let $G_3 = \{g_3 \, . \, x \in]0,L[\mapsto g_3(x) = \partial_3 h, \; h \in H_0^1(]0,L[)\}$, then, if the force component $g_3 \in G_3 \cap H_0^{2N-1}(]0,L[)$ for $N \geq 1$, there exist constants C and K, independent of ε, δ and N, such that :*

$$\|u^\delta(\varepsilon) - u_N^\delta(\varepsilon)\|_{H^1(\Omega^{\delta,1})} \leq CNK^{N-1}\varepsilon^{2N-2},$$
$$\|u^{\delta,\varepsilon} - u_N^{\delta,\varepsilon}\|_{H^1(\Omega^{\delta,\varepsilon})} \leq CNK^{N-1}\varepsilon^{2N-1}.$$

The approximation space $V_N(\Omega^{\delta,1})$ is given by :

$$V_N(\Omega^{\delta,1}) = \{v = (v_i) \in V(\Omega^{\delta,1}) :$$
$$v_\alpha = \sum_{k=0}^{N} P_\alpha^{k\delta}(x_1,x_2) v_\alpha^k(x_3), v_\alpha^k \in H_0^1(]0,L[), P_\alpha^{k\delta} \in H^1(\omega^{\delta,1}),$$
$$v_3 = \sum_{k=0}^{N} Q^{k\delta}(x_1,x_2) v_3^k(x_3), v_3^k \in H_0^1(]0,L[), Q^{k\delta} \in H^1(\omega^{\delta,1}),$$
$$\forall 0 \leq k \leq N, \quad (\textit{no sum on } \delta) \quad \}$$

and the basis functions are defined, by recurrence, as the solution to the following boundary value problems :

$$\begin{cases} \int_{\omega^\delta} [\lambda e_{\alpha\alpha}(P^{k\delta}) e_{\beta\beta}(s) + 2\mu e_{\alpha\beta}(P^{k\delta}) \partial_\alpha s_\beta] d\omega^{\delta,1} + \\ + \int_{\omega^\delta} \lambda Q^{(k-1)\delta} e_{\alpha\alpha}(s) - \\ - \mu(\partial_\alpha Q^{(k-1)\delta} + P_\alpha^{(k-1)\delta}) s_\alpha d\omega^{\delta,1} = 0, \quad \forall s \in [H^1(\omega^{\delta,1})]^2; \end{cases}$$

$$\begin{cases} \int_{\omega^\delta} \mu(\partial_\alpha Q^{k\delta} + P_\alpha) \partial_\alpha r d\omega^{\delta,1} - \\ - \int_{\omega^\delta} [(\lambda + 2\mu) Q^{(k-1)\delta} + \lambda \partial_\alpha P_\alpha^{k\delta}] r d\omega^{\delta,1} = \delta_{k1} \int_\gamma r d\gamma, \\ \forall r \in H^1(\omega^{\delta,1}). \end{cases}$$

with $P_\alpha^{-1\delta} = 0$ *and* $Q^{-1\delta} = 0$. □

IV. THE HOMOGENIZATION RESULT

We proceed in order to study the behaviour of approximations $u_N^\delta(\varepsilon)$, as δ becomes very small. The following result holds.

Theorem 2. *Let* $P^{k\delta}$ *and* $Q^{k\delta}$ *be the basis functions defined in Theorem 1. For each* k *there exist extensions* $\tilde{P}^{k\delta}$ *and* $\tilde{Q}^{k\delta}$ *of* $P^{k\delta}$ *and* $Q^{k\delta}$, *respectively, that are bounded, in* $[H^1(\omega)]^2$ *and* $H^1(\omega)$, *respectively, independently of* δ. *Any other extensions* $\tilde{P}^{k\delta}$ *and* $\tilde{Q}^{k\delta}$, *bounded independently of* δ, *satisfy the following convergences*

$$\tilde{P}^{k\delta} \to P^*, \quad \text{weakly in } [H^1(\omega)]^2,$$
$$\tilde{Q}^{k\delta} \to Q^*, \quad \text{weakly in } H^1(\omega),$$

where P^{k*} *and* Q^{k*} *are uniquely defined by the following recurrence formulas :*

$$P^{-1*} = 0, \quad Q^{-1*} = 0, \quad (k = -1);$$
$$P^{k*} = \overline{P}^{k*} + {}^R P^{k*}, \quad Q^{k*} = \overline{Q}^{k*} + {}^R Q^{k*}, \quad (k \geq 0);$$

and where $\overline{P}^{k*}$ *is the solution of the following plane deformation elasticity problem in* ω :

$$\begin{cases} -\partial_\alpha S_{\alpha\beta}^{k*} = \mu a_{\alpha\beta}^* (P_\alpha^{(k-1)*} + \partial_\alpha Q^{(k-1)*}) + \\ + \lambda(\Theta \delta_{\alpha\beta} + s_{\rho\rho\alpha\beta}^*) \partial_\alpha Q^{(k-1)*}, \quad \text{in } \omega, \\ S_{\alpha\beta}^{k*} n_\alpha = -\lambda(\Theta \delta_{\alpha\beta} + s_{\rho\rho\alpha\beta}^*) Q^{(k-1)*} n_\alpha, \quad \text{on } \gamma, \\ \int_\omega \overline{P}_\alpha^{k*} d\omega = \int_\omega (\overline{P}_1^{k*} x_2 - \overline{P}_2^{k*} x_1) d\omega = 0, \end{cases}$$

with $S_{\alpha\beta}^{k*}$ *given by*

$$S_{\alpha\beta}^{k*} = S_{\alpha\beta\zeta\rho}^* e_{\zeta\rho}(\overline{P}^{k*}), \quad S_{\alpha\beta\zeta\rho}^* = \Theta S_{\alpha\beta\zeta\rho} + S_{\alpha\beta\tau\sigma} s_{\tau\sigma\zeta\rho}^*,$$
$$s_{\alpha\beta\zeta\rho}^* = \int_{Y^*} e_{\alpha\beta}(\chi^{\zeta\rho}) dY^*,$$

for $\chi^{\zeta\rho} = (\chi_\alpha^{\zeta\rho})$ *defined, up to a constant vector, as the solution of the following problem, in the unit cell* Y^* :

$$\begin{cases} \chi^{\zeta\rho} \in [H^1(Y^*)]^2, \quad \chi^{\zeta\rho}\ Y^*\text{-periodic}, \\ \int_{Y^*} [2\mu e_{\alpha\beta}(\chi^{\zeta\rho}) + \lambda e_{\rho\rho}(\chi^{\zeta\rho}) \delta_{\alpha\beta}] e_{\alpha\beta}(\varphi) dY^* = \\ = -\int_{Y^*} [2\mu e_{\zeta\rho}(\varphi) + \lambda e_{\rho\rho}(\varphi) \delta_{\zeta\rho}] dY^*, \end{cases}$$

the elasticity tensor $(S_{\alpha\beta\zeta\rho}^*)$ *being positive definite, satisfying* $S_{\alpha\beta\zeta\rho}^* = S_{\beta\alpha\zeta\rho}^* = S_{\zeta\rho\alpha\beta}^*$, $A^* = (a_{\alpha\beta}^*)$ *is a* 2×2 *symmetric and positive definite matrix, whose coefficients are defined by* $a_{\alpha\beta}^* = \Theta \delta_{\alpha\beta} + \int_{Y^*} \partial_\alpha \chi_\beta = \int_{Y^*} \nabla(\chi_\alpha + y_\alpha) \nabla(\chi_\beta + y_\beta)$, *and where* χ_β *is the unique solution of the following problem defined in the cell* Y^* :

$$\begin{cases} \chi_\beta\ Y^*\text{-periodic}, \quad \int_{Y^*} \chi_\beta = 0, \\ -\Delta \chi_\beta = 0, \quad \text{in } Y^*, \\ \partial_n \chi_\beta = -n_\beta, \quad \text{on } \partial T. \end{cases}$$

Moreover, $\overline{Q}^{k*}$ *is the solution of the following membrane or (anisotropic) torsion type problem in linearized elasticity, for a two-dimensional body occupying volume* ω :

$$\begin{cases} -\mu\, div(A^* \nabla \overline{Q}^{k*}) = \\ = \mu\, div(A^* \overline{P}^{k*}) + \lambda[\Theta e_{\beta\beta}(\overline{P}^{k*}) + s_{\beta\beta\zeta\rho}^* e_{\zeta\rho}(\overline{P}^{k*})] + \\ + \frac{\lambda^2}{2(\lambda+\mu)} s_{\rho\rho\beta\beta}^* Q^{(k-1)*} + (\lambda + 2\mu) \Theta Q^{(k-1)*}, \quad \text{in } \omega, \\ (A^* \nabla \overline{Q}^{k*}) \cdot n = -(A^* \overline{P}^{k*}) \cdot n + \frac{1}{\mu} \delta_{k1}, \quad \text{on } \gamma, \\ \int_\omega Q^{k*} d\omega = 0; \end{cases}$$

with

$${}^R P_1^{k*} = a_1^{k*} + b^{k*} x_2, \quad {}^R P_2^{k*} = a_2^{k*} + b^{k*} x_2,$$
$${}^R Q^{k*} = -b^{k*} w^* - a_1^{k*} x_1 - a_2^{k*} x_2 + c^{k*},$$

where w^* *is the homogenized solution of Saint Venant's torsion function and where constants* a_α^{k*}, b^{k*} *and* c^{k*} *depend only on the geometry of the cross section.* □

Theorem 3. *Let* $f^\varepsilon \in [L^2(\Omega^\varepsilon)]^3$ *and* $g^\varepsilon \in [L^2(\Gamma_1^\varepsilon)]^3$. *The solution* $u^{\delta,\varepsilon} \in V^{\delta,\varepsilon}$ *of the three-dimensional elasticity problem, where the volumic forces are the restrictions of* f^ε *to* $\Omega^{\delta,\varepsilon}$ *and where the surface forces coincide in* Γ_1^ε *with* g^ε, *admits an extension in* $[H^1(\Omega^\varepsilon)]^3$, *bounded independently of* δ. *If* $\tilde{u}^{\delta,\varepsilon}$ *is any extension of* $u^{\delta,\varepsilon}$; *bounded in* $[H^1(\Omega^\varepsilon)]^3$, *independently of* δ, *then, for each fixed* ε *and as* $\delta \to 0$, $\tilde{u}^{\delta,\varepsilon}$ *converges to* $u^{*,\delta}$, *weakly in* $[H^1(\Omega^\varepsilon)]^3$, *where* $u^{*,\delta}$ *is the unique solution of the following problem :*

$$\begin{cases} u^{*,\delta} \in V^\varepsilon = \{v = (v_i) \in [H^1(\Omega^\varepsilon)]^3 : v = 0 \quad \text{on } \Gamma_0^\varepsilon\} \\ \int_{\Omega^\varepsilon} \sigma_{ij}^{*,\varepsilon} e_{ij}^\varepsilon(v) d\Omega^\varepsilon = \\ \quad = \Theta \int_{\Omega^\varepsilon} f_i^\varepsilon v_i d\Omega^\varepsilon + \int_{\Gamma_1^\varepsilon} g_i^\varepsilon v_i d\Gamma_1^\varepsilon, \quad \forall v \in V^\varepsilon, \\ \sigma_{ij}^{*,\varepsilon} = 2\mu[\Theta e_{ij}^\varepsilon(u^{*,\varepsilon}) + q_{ijkl}^* e_{kl}^\varepsilon(u^{*,\varepsilon})] + \\ \quad + \lambda[\Theta e_{pp}^\varepsilon(u^{*,\varepsilon}) + q_{ppkl}^* e_{kl}^\varepsilon(u^{*,\varepsilon})] \delta_{ij}, \end{cases}$$

where the coefficients $qijkl^*$ *satisfy :*

$$q_{\alpha\beta\zeta\rho}^* = s_{\alpha\beta\zeta\rho}^*, \quad q_{\alpha\beta33}^* = \frac{\lambda}{2\mu} s_{\rho\rho\alpha\beta}^* - \frac{\lambda^2}{4\mu(\lambda+\mu)} s_{\rho\rho\zeta\zeta}^* \delta_{\alpha\beta},$$
$$q_{3\alpha3\beta}^* = q_{3\beta3\alpha}^* = q_{\alpha33\beta}^* = q_{\alpha\beta33}^* = \frac{1}{2}(a_{\alpha\beta}^* - \Theta \delta_{\alpha\beta}),$$

and all the others vanish. □

Theorem 4. *Under the same hypothesis as before and using the previous notations, there exist constants* $\overline{C}$ *and* $\overline{K}$, C^* *and* K^*, *independent of* ε, δ *and* N, *such that*

$$\|u^*(\varepsilon) - u_N(\varepsilon)\|_{H^1(\Omega)} \leq \overline{C} N \overline{K}^{N-1} \varepsilon^{2N-2},$$
$$\|u^{*,\varepsilon} - u_N^\varepsilon\|_{H^1(\Omega^\varepsilon)} \leq \overline{C} N \overline{K}^{N-1} \varepsilon^{2N-1},$$
$$\|u^*(\varepsilon) - u_N^*(\varepsilon)\|_{H^1(\Omega)} \leq C^* N (K^*)^{N-1} \varepsilon^{2N-2},$$
$$\|u^{*,\varepsilon} - u_N^{*\varepsilon}\|_{H^1(\Omega^\varepsilon)} \leq C^* N (K^*)^{N-1} \varepsilon^{2N-1}.$$

□

SINGULAR ELEMENTS FOR 3D CRACK PROBLEMS

L.GAVETE*, F.MICHAVILA*, U.KINDELAN* and D.GARCIA VERA**

(*) Departamento de Matemática Aplicada y Métodos Informáticos. E.T.S.I. Minas. Universidad Politécnica de Madrid. C/ Ríos Rosas, 21 28003 Madrid (Spain)

(**) Departamento de Matemática Aplicada. E.T.S.I. Industriales. Universidad de Las Palmas de Gran Canaria. Islas Canarias (Spain)

Abstract.- In any adaptation of a numerical technique for treating a singular situation, it is advantageous to have the knowledge of the form of the singularity. In this paper we extend the quadratic mapping technique to the treatment of singularities in three dimensional problems. Element types are developed for a three dimensional situation and comparison of results is made for a plate with circumferential crack. With the quadratic mapping technique we can perform fracture mechanics computations with rather simple finite element meshes. The element is shown to be able to reproduce a singular strain field in three dimensions for a curved singularity line.

I. INTRODUCTION

The analysis of the finite element method usually relies on the assumption that the solution of the given problem is regular enough. However, the implementation of the method is very often done on problems with polygonal domains which prevent the solution from being smooth in some points or lines. According Grisvard [1] the presence of corners lead to the singular behaviour of the solution only near the corners. This singular behaviour occurs even when the data of the problem are very smooth. It strongly affects the accuracy of the finite element method troughout the whole domain. A considerable body of analysis now exists showing that singularities can occur at such boundary points and lines, with the effect that the regularity of the solution is reduced from what is expected for such problems when the regions have smooth boundaries. However, many problems in potential theory and linear elasticity occur in regions which contain sharp corners and edges.

In any adaptation of a numerical technique for treating a singular situation, it is advantageous to have the knowledge of the form of the singularity. The form of the singularity is determined by the combination of geometry and boundary conditions. If we know the form of a singularity in general non convex domain in two dimensions, we can use the technique of Aalto [2] for locating the nodal points of standard isoparametric quadrilateral elements, properly around the singularity. The use of this quadratic mapped element for treating singularities has been shown by Michavila, Gavete, Díez and Whiteman [3,4,5]. Also, it has been demonstrated by Gavete, Michavila and Díez [6,7] that the Serendipity and Lagrange quadratic mapped elements can be applied in linear elastic fracture, because their strains and displacements in the vicinity of the crack tip, are appropiate to the form of the singularity.

In this paper we extend the quadratic mapping technique to the three dimensional problems. Element types are developed for a three dimensional situation and comparison of the results is made for a plate with circumferential crack. With the quadratic mapping technique we can perform fracture mechanics computations with rather simple finite element meshes.

The element is shown to be able to reproduce a singular strain field in three dimensions for a circumferential singularity line. Other forms of singularity lines could be possible.

II A NEW APPROACH TO THE TREATMENT OF 3-D CRACK PROBLEMS.

Our starting point is the work of Aalto, Gavete, Michavila and Díez [2-7], and in this paper we calculate a new mapping, giving a special singular finite element, which is used to approach a 3-D circumferential crack problem. The strain form approach for this new singular element is appropiate to the singularities involved, as it is demonstrated in this paper. The mapping is as follows:

$$x = \frac{a}{4}(\lambda^2-\mu^2)+\left[(r+a)-\frac{a}{4}(\lambda^2-\mu^2)\right]\frac{(1-\cos\nu\alpha)}{2}$$

$$y = \frac{a}{2}\lambda\mu \qquad (1)$$

$$z = \left[(r+a)-\frac{a}{4}(\lambda^2-\mu^2)\right]\operatorname{sen}\frac{\nu\alpha}{2}$$

being $\lambda, \mu, \nu \in [0,2]$

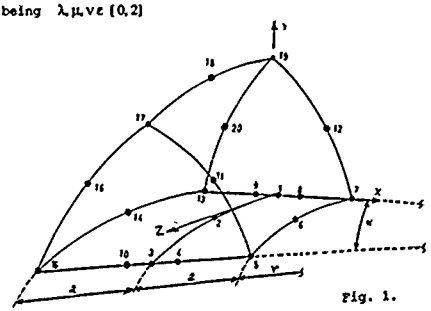


Fig. 1.

This mapping is applied to the standard 20-nodes brick finite element giving the element of the fig. 1 in physical space. The singularity line (crack line) is placed along 1 - 2 - 3.

The approximations of the strains ε_{11}, ε_{22} and ε_{12} in planes ortogonal to the crack line are

$$E_{11}=\frac{\partial U(\lambda,\mu,\nu)}{\partial x}=\frac{\partial U}{\partial\lambda}\frac{\partial\lambda}{\partial x}+\frac{\partial U}{\partial\mu}\frac{\partial\mu}{\partial x}+\frac{\partial U}{\partial\nu}\frac{\partial\nu}{\partial x}$$

$$E_{22}=\frac{\partial V(\lambda,\mu,\nu)}{\partial y}=\frac{\partial V}{\partial\lambda}\frac{\partial\lambda}{\partial y}+\frac{\partial V}{\partial\mu}\frac{\partial\mu}{\partial y}+\frac{\partial V}{\partial\nu}\frac{\partial\nu}{\partial y} \qquad (2)$$

$$E_{12}=\frac{1}{2}\left[\frac{\partial U(\lambda,\mu,\nu)}{\partial y}+\frac{\partial V(\lambda,\mu,\nu)}{\partial x}\right]$$

For the mapping (1) we have

$$|J|=\left|\frac{\partial(x\,y\,z)}{\partial(\lambda\mu\nu)}\right|=\frac{\alpha^2 c}{8}\left[(r+a)-\frac{a}{4}(\lambda^2+\mu^2)\right](\lambda^2+\mu^2) \qquad (3)$$

The forms of $U(\lambda,\mu,\nu)$, $V(\lambda,\mu,\nu)$ are

$$U(\lambda,\mu,\nu)=\alpha_1+\alpha_2\lambda+\alpha_3\mu+\alpha_4\nu+\alpha_5\lambda^2+\alpha_6\mu^2+\alpha_7\nu^2$$
$$+\alpha_8\lambda\mu+\alpha_9\lambda\nu+\alpha_{10}\mu\nu+\alpha_{11}\lambda^2\mu+\alpha_{12}\lambda\mu^2+\alpha_{13}\mu^2\nu+\alpha_{14}\mu\nu^2+$$
$$+\alpha_{15}\lambda^2\nu+\alpha_{16}\lambda\nu^2+\alpha_{17}\lambda\mu\nu+\alpha_{18}\lambda^2\mu\nu+\alpha_{19}\lambda\mu^2\nu+\alpha_{20}\lambda\mu\nu^2 \qquad (4)$$

and similarly for $V(\lambda,\mu,\nu)$ changing α_i by β_i $(i=1,\ldots,20)$

$$E_{11}=\left[\frac{\partial U}{\partial\lambda}\frac{a\lambda\alpha}{4}\cos\frac{\nu\alpha}{2}\left[(r+a)-\frac{a}{4}(\lambda^2+\mu^2)\right]\right.$$
$$+\frac{\partial U}{\partial\mu}\frac{a\alpha\mu}{4}\left[(r+a)-\frac{a}{4}(\lambda^2+\mu^2)\right]$$
$$-\frac{\partial U}{\partial\nu}\left[(r+a)-\frac{a}{4}(\lambda^2+\mu^2)\right]$$
$$\left.\left[\frac{a\alpha\lambda}{4}\,\mathrm{sen}\,\frac{\nu\alpha}{2}\right]\right]\Big/|J| \qquad (5)$$

and for $\lambda=0$

$$E_{11}=\frac{\alpha_3+2\alpha_4\mu+\alpha_{10}\nu+2\alpha_{13}\mu\nu+\alpha_{14}\nu^2}{\frac{a}{2}\mu} \qquad (6)$$

and hence for $\nu=$ cte, i.e. in a plane ortogonal to the crack line

$$E_{11}=\frac{\widetilde{A}_1+\widetilde{A}_2}{\mu}=A_1+\frac{A_2}{\sqrt{r}} \qquad (7)$$

where, $\widetilde{A}_i$ and A_i $(i=1,2)$ are constants.

Similar results we obtain for ($\mu=0$, $\nu=$constant) and for ($\lambda=K\mu$, $\nu=$ constant), radial lines emanating from the singularity in a plane ortogonal to the crack line (1 - 2 - 3 of fig. 1), where $K>0$ is a constant. Also we obtain similar results for E_{22} and E_{12} ($\nu=$ constant). Thus for small r the $r^{-1/2}$ term dominates in all directions emanating from the line of singularity in planes ortogonal to this line, giving the approximations to the gradients the correct "$r^{-1/2}$" singular form as required by the true solution. Similar results to those obtained in the case of mapping the 20-nodes brick element, can be obtained by mapping the 27-nodes quadratic isoparametric Lagrange brick element in a similar way.

III. AN EXAMPLE CONTAINING A SINGULARITY LINE

The 3-Dimension crack models represent a very important problem nowadays in mechanical engineering, considering that pipe and pressure vessel cracks can be detected, before the broken pressure arrives. Width and length intervene very directly in this models, and even the crack shape is important.

As an example we have studied the problem of a plate with a circumferential crack which is under an uniform stress field.

Figure 2 shows plate, subject to an uniform stress field and having a circular center crack. Data are a/t= 0.4, L/t= 2.5 and a/w= 0.2.

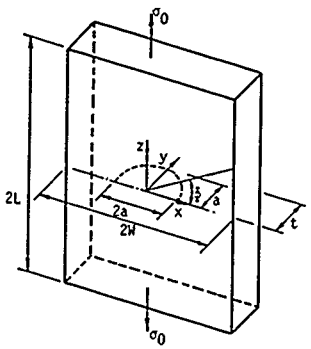


Fig. 2

In order to use the new mapping defined in (1) for the case of circumferential crack line, fig.3 model, has been carried out. After the displacements have been obtained, we use the Manu's formula [8], to calculate the SIF.

As a difference with the quarter point elements, here we don't have negative pivots during the factorization which is carried out in the equations system solution or other numerical difficulties as reported by Peano and Pasini in [9]. The results obtained for S.I.F. versus the orientation (Φ angle of figure 1) are shown in figure 4.

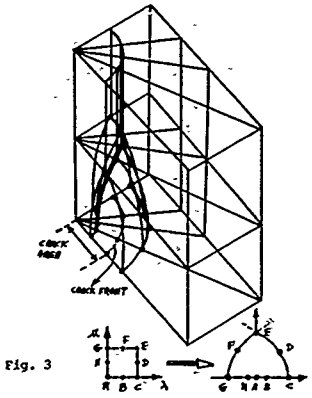


Fig. 3

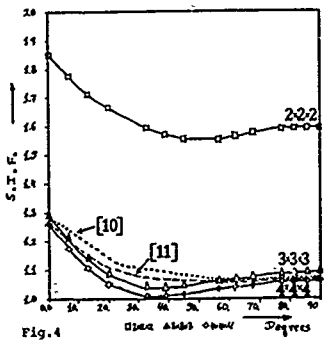


Fig.4

The performed errors when compared with reforences [10] and [11] can reach up to 5% and even lower, which represents an acceptable error level, if we consider that reference models of Miyazaki, Watanabe and Yagawa [10] and Yagawa, Ichimiya and Ando [11] have more finite elements. In general , these errors depend on the numerical integration order, on the diameter and on the elements aspect ratio of the model.

Also a very good agreement has been obtained with the accurate results of Schnack and Karaosmanoglu [12].

IV. CONCLUSIONS

We have compared ours curves of the SIF values versus the angle Ø, with those obtained by Miyazaki, Watanabe, Ichimiya, Ando and Yagawa [10-11] , when they solve a similar example with different strategies. Their results are very much the same than ours (integration orders 3 and 4), but our grid is much less refinate. This method has the advantage of its simplicity, it is possible to solve problems with much less refinate grids. We have demonstrated that a new singular element can approach the behaviour of curved cracks in 3-D, calculating with high precission the stress intensity factor. Others forms of singularity lines could be possible. It would be convenient also to study the influence of the geometric shape of the crack.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish Commission of Technical Scientifical Research (CICYT), Project (PA 85-0053) and The Canary Islands Research Proyect ("Modelización de singularidades mediante elementos finitos Hostern. Aplicaciones al caso de materiales compuestos 1991-92"). The writers gratefully acknowledge this support.

REFERENCES

1 *Grisvard, P.:* Elliptic problems in nonsmooth domains. (Monographs and Studes in Mathematics, 24). Ed. Pitman, (1985).

2 *Aalto, J.:* Singularity elements for seepage problems. Int. J. Anal. Meth. Geomech. Vol. 9, 185-196, (1985).

3 *Gaveto, L., Michavila, F., Díez, F., and Whiteman, J. R.* Generalization of the mapping technique of Aalto for producing finite elements with singularities. MAFELAP 1987. (Ed. J.R. Whiteman) pp. 541-553. London. Academic Press (1988).

4 *Díez, F., Gavete, L. and Michavila, F.* Nuevas Técnicas de tratamiento de singularidades en campos escalares y vectoriales. III Encuentro del Grupo Español de Fractura. Sigüenza (1986) (in spanish).

5 *Michavila, F., Gavete, L. and Díez, F.* Two different approachs for the treatment of boundary singularities, Numerical Methods for partial differential equations, 255-282, (1988).

6 *Gavete, L., Michavila, F. and Díez, F.* A new singular finite element in linear elasticity. Computational Mechanics 4, 361-367 (1989).

7 *Gavete, L., Michavila, F.,and Díez, F.* Análisis de gradientes en elementos singulares cuadráticos de Serendepity y de Lagrange. Revista Internacional de Métodos Numéricos para cálculo y diseño en Ingeniería. Vol. 3, 2, 153-171, (1987) (in spanish).

8 *Hanu, C.* Pure opening mode stress-intensity factor computation for elliptical crack fronts. Int. J. for Num.Meth. Eng. Vol. 21, 1547-1553, (1985).

9 *Peano, A. and Pasini, A.* A warning against misuse of quarter point elements. Int. J. Num. Meth. Eng., 314-319 (1982)

10 *Miyazaki, N., Watanabe, T. Yagawa, G.* Calculation of stress intensity factors of surface cracks in complex structures: application of efficient computer program EPAS-J1. Nuclear Engineering and Design 68, pag. 71-85 (1981).

11 *Yagawa, G., Ichimiya, M. and Ando, Y.* Two and three dimensional analysis of stress intensity factors based on discretization error in finite elements. Numerical Methods in fracture mechanics, 249-267. Ed. Luxmoore and Owen. Pineridge Press (1978).

12 *Schnack E. and Karaosmanoglu.*. Unpublished paper. Institute of Solid Mechanics. Karlsruhe University, (1989).

COEFFICIENT FORMULAE FOR ASYMPTOTIC EXPANSIONS OF SOLUTIONS OF ELLIPTIC BOUNDARY VALUE PROBLEMS NEAR CONICAL POINTS

ANNA-MARGARETE SÄNDIG
University of Rostock
Department of Mathematics
Universitätsplatz
O-2500 Rostock
Germany

Abstract

It is well known that singularities are present in solutions of elliptic boundary value problems in domains with conical boundary points. The solution consists of singular terms, which appear in a neighbourhood of a conical point, and a more regular term. The coefficients of the singular terms, the so-called stress intensity factors, are especially of interest for applications We describe a method, how some of them may be calculated, if the right hand sides are from standard Sobolev spaces. In some cases the coefficients are unstable and a stabilization procedure is necessary. We handle as examples boundary value problems for the Laplace equation in two and three dimensional domains.

I. THE ASYMPTOTIC EXPANSION

Let Ω be an open subset of $\mathbb{R}^n$, whose $n-1$ dimensional boundary $\partial\Omega$ is smooth with exception of one conical point 0. We consider an elliptic boundary value problem with constant coefficients

$$A(D_x)u(x) = \sum_{|\alpha|\le 2m} a_\alpha D_x^\alpha u(x) = f(x) \text{ in } \Omega \qquad (1)$$

$$B_j(D_x)u(x) = \sum_{|\alpha|\le m_j} b_{j\alpha} D_x^\alpha u(x) = g_j(x) \text{ on } \partial\Omega \setminus 0, j = 1,\dots,m.$$

The right hand sides are from the Sobolev spaces $W^{k,p}(\Omega)$ and the trace spaces $W^{k+2m-m_j-\frac{1}{p},p}(\partial\Omega)$. If $n = 2$, the boundary conditions are also considered on pieces of the boundary. In this case we assume, that the functions $g_j(x)$ satisfy compability conditions. Since the right hand sides can be splitted into Taylor polynomials and functions from certain weighted Sobolev spaces, we get e g. the following asymptotic expansion of a weak solution u from $W^{2,p}(\Omega)$:

$$u = \eta(r)\left(\sum_{l=0}^{k+2m-s-1} r^l v_l + \sum_{Re\alpha_\nu \in I} r^{\alpha_\nu} u_{\alpha_\nu}\right) + w, \qquad (2)$$

where

$$v_l = \sum_{i=0}^{M_l} (\ln r)^i \psi_i(l,\omega), \qquad (3)$$

$$u_{\alpha_\nu} = \sum_{i=0}^{M_{\alpha_\nu}} C_{\alpha_\nu i} \sum_{j=0}^{i} (\ln r)^j e_{i-j}(\alpha_\nu,\omega), \qquad (4)$$

$I = \left(-\frac{n}{2}+m, -\frac{n}{p}+k+2m\right)$, and $w \in W^{k+2m,p}(\Omega)$ for $p \neq 2$.

We denote by (r,ω) the spherical coordinates, $s = s(n,p)$ is an integer. $M_l > 0$ is the multiplicity of the "eigenvalue" l of a generalized eigenvalue problem, which we get from (1) introducing (r,ω) and using the Mellin transform with respect to r. $M_l = 0$ means, that l is not an eigenvalue. In the second term the complex numbers α_ν are eigenvalues of the multiplicity M_{α_ν} and $M'_{\alpha_\nu} = M_{\alpha_\nu} - 1$.

The first term of (2) comes from the polynomial parts of the right hand sides [1], the second term is known from the theory in weighted Sobolev spaces [1], [2].

II. THE PLANE CASE

Due to the compatibility conditions we restrict to the problem

$$Au = 0 \text{ in } \Omega \qquad (5)$$

$$B_j^{(q)}u = g_j^{(q)} \text{ on } \Gamma_q, \quad \bigcup_{q=1}^{Q} \overline{\Gamma_q} = \partial\Omega, \; j = 1,\dots,m.$$

Assume that the domain Ω coincides in a neighbourhood of a corner point $\overline{\Gamma_q} \cap \overline{\Gamma_{q+1}} = O_q = O$ with the infinite cone $K = \{(r,\omega), 0 < r < \infty, \omega_0^+ = 0 < \omega < \omega_0 = \omega_0^-\}$ with the sides $\Gamma^\pm$. Introducing polar coordinates we write the differential operators as

$$A(D_x) = r^{-2m} L(rD_r, \omega, D_\omega)$$

$$B_j^\pm(D_x) = r^{-m_j^\pm} M_j^\pm(r, Dr, \omega, D_\omega)\Big|_{\Gamma^\pm}.$$

We say, the complex number α is an eigenvalue of the operator $\mathcal{A}(\lambda) = \left\{L(\lambda,\omega,D_\omega), M_j^\pm(\lambda,\omega,D_\omega)\right\}$ if there is a nontrivial solution $e_0(\lambda,\omega)$ of

$$\mathcal{A}(\lambda)e(\lambda,\omega) = 0.$$

We now split the right hand sides of (5) :

$$g_j^\pm(r,\omega_0^\pm) = \sum_{l=0}^{k+2m-s-1} G_{l,j}^\pm(0,\omega_0^\pm) r^l + g_j^\pm(r,\omega_0^\pm),$$

where $G_{l,j}^\pm(0,\omega_0^\pm) = \frac{1}{l!}\frac{\partial^l g_j^\pm}{\partial r^l}(0,\omega_0^\pm)$,

$s = 0$ for $2 < p$ and $s = 1$ for $1 < p < 2$.

The functions v_l in the expansion (2) are solutions of

$$Ar^l v = 0 \quad \text{in } K$$

$$B_j^\pm r^l v = G_{l-m_j^\pm,j}^\pm(0,\omega_0^\pm) r^{l-m_j^\pm} \text{ on } \Gamma^\pm.$$

They can be calculated "easily", starting from the general solution of the ordinary differential equation $L(\lambda,\omega,D_\omega)e(l,\omega) = 0$ and using the ansatz (3).

Thus we get for the Dirichlet problem for the Laplace equation

$$v_l = \begin{cases} G_l(0,0)\left\{\cos l\omega + \frac{1-\cos l\omega_0}{\sin l\omega_0}\sin l\omega\right\} & \text{for } l\omega_0 \neq \nu\pi \\ G_l(0,0)\left\{\cos l\omega + \frac{1-\cos l\omega_0}{\omega_0 \cos l\omega_0}(\ln r \sin l\omega + \omega\cos l\omega)\right\} & \text{for } l\omega_0 = \nu\pi \end{cases} \qquad (6)$$

It is evident that one coefficient in the first row of v_l is unbounded (unstable), if ω_0 is from a neighbourhood of $\frac{\nu\pi}{l}$. This behavior influences also the coefficients c_{α_ν} of u_{α_ν} (in our example is $\alpha_\nu = \frac{\nu\pi}{\omega_0}$, $u_{\alpha_\nu} = c_{\alpha_\nu}\sin\alpha_\nu\omega$). Following an idea of V.G.Maz'ya [3] we get a stable asymptotics, organizing the sums of (2) as follows (here for our example): In a neighbourhood of a critical angle $\frac{\nu\pi}{l}$ we write

$$r^l G_l(0,0)\left(\frac{1-\cos l\omega_0}{\sin l\omega_0}\right)\sin l\omega + r^{\frac{\pi\pi}{\omega_0}} c_{\alpha_\nu}\sin\alpha_\nu\omega$$

$$= G_l(0,0)(1-\cos l\omega_0)\left[\frac{r^l\sin l\omega - r^{\frac{\pi\pi}{\omega_0}}\sin\alpha_\nu\omega}{\sin l\omega_0}\right]$$

$$+ r^{\frac{\pi\pi}{\omega_0}}\left[c_{\alpha_\nu} + \frac{G_l(0,0)(1-\cos l\omega_0)}{\sin l\omega_0}\right]\sin\alpha_\nu\omega \qquad (7)$$

The new coefficients $c_l = G_l(0,0)(1-\cos l\omega_0)$ and $c'_{\alpha_\nu} = c_{\alpha_\nu} + \frac{G_l(0,0)(1-\cos l\omega_0)}{\sin l\omega_0}$ are bounded, if ω_0 is from a neighbourhood of $\frac{\pi\pi}{l}$ and the first term of (7) converges for $\omega_0 \to \frac{\pi\pi}{l}$ to the term of the second row of (6).
A general stabilization procedure is given in [3] and [4].

III. THE THREE DIMENSIONAL CASE

We consider the problem

$$Au = 0 \quad \text{in } \Omega$$
$$B_j u = g_j \quad \text{on } \partial\Omega|0,$$

where the domain Ω coincides in a neighbourhood of 0 with a circle cone $K = \{(r,\varphi,\vartheta) = (r,\underline{\omega}) : 0 < r < \infty,\ 0 \le \varphi < 2\pi,$ $0 < \vartheta < \vartheta_0\}$. Analogously to the plane case we define $\mathcal{A}(\lambda)$ and consider a decomposition of g_j:

$$g_j(r,\varphi,\vartheta_0) = \sum_{l=0}^{k+2m-m_j-s-1} r^l G_{jl}(0,\varphi,\vartheta_0) + \tilde{g}_j(r,\varphi,\vartheta_0),$$

where

$$s = \begin{cases} 0 & \text{for } p > 3 \\ 1 & \text{for } \frac{3}{2} < p < 3 \\ 2 & \text{for } 1 < p < \frac{3}{2} \end{cases}$$

and

$$G_{jl}(0,\varphi,\vartheta_0) = \frac{\partial^l g_j(0,\varphi,\vartheta_0)}{\partial r^l}\frac{1}{l!}$$

It is meaningful to assume that $g_j(r,\varphi,\vartheta_0) = g_j(r,\varphi+2\pi,\vartheta_0)$ and to consider a Fourier expansion of G_{jl}:

$$G_{jl}(0,\varphi,\vartheta_0) = \sum_{h=0}^{\infty} A_h^l(l,\vartheta_0)\cos h\varphi + B_h^l(l,\vartheta_0)\sin h\varphi$$

The functions v_l of (2), which satisfy the equations

$$L(l,\omega,D_\omega)v_l = 0 \quad \text{in } K$$
$$M_j(l,\omega,D_\omega)v_l = G_{jl-m_j}(0,\varphi,\vartheta_0)$$

can be calculated, writing v_l as Fourier series with respect to φ. Thus we get for the Neumann problem for the Laplace equation (the index j is cancelled)

$$v_l = \sum_{h=0}^{\infty}\left(\frac{-A_h(l-1,\vartheta_0)}{\frac{\partial}{\partial\vartheta}P_l^{-h}(\cos\vartheta_0)}P_l^{-h}(\cos\vartheta)\cos h\varphi\right.$$

$$\left. -\frac{B_h(l-1,\vartheta_0)}{\frac{\partial}{\partial\vartheta}P_l^{-h}(\cos\vartheta_0)}P_l^{-h}(\cos\vartheta)\sin h\varphi\right),$$

if $\frac{\partial}{\partial\vartheta}P_l^{-h}(\cos\vartheta_0) \neq 0$ for all h. $P_l^{-h}(\cos\vartheta)$ are Legendre functions of the first kind.

If there is for a given ϑ_0 an index h_0 with $\frac{\partial}{\partial\vartheta}P_l^{-h_0}(\cos\vartheta_0) = 0$ (see Figure), then l is an eigenvalue and terms with $\ln r$ as in (3) occur.
The asymptotic expansion is unstable in this case and we can apply a stabilization procedure analogously to the plane case.

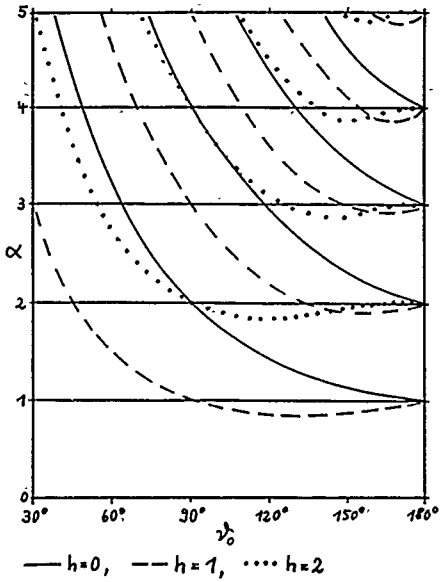
References

[1] Kondrat'ev, V. A. Boundary value problems for elliptic equations on domains with conical or angular points. Trudy Moskov. Mat. Obshch. 16 (1967) 209-292

[2] Maz'ya, V. G., Plamenevskij, B. A. On the coefficients in the asymptotics of solutions of elliptic boundary value problems with conical points. Math. Nachr. 76 (1977), 29-60

[3] Maz'ya, V.G., Roßmann J. On a problem of Babushka, preprint University Linköpping LiTH-Mat-R-90-33

[4] Sändig, A-M., Möller, K. Coefficient formulae for asymptotic expansions of solutions of elliptic boundary value problems near conical boundary points, to appear



Singularities of Three - Dimensional Elastic Fields

K. Volk
IBM Scientific Center
Institute for Supercomputing and Applied Mathematics
D - 6900 Heidelberg, Germany

Abstract. The computation of the singular behavior of an elastic field near a three - dimensional vertex subject to displacement boundary conditions leads to a one - dimensional integral equation on a piecewise circular curve γ in R^3 depending holomorphically on a complex parameter. The corresponding spectral points and packets of generalized eigenfunctions characterize the desired singular behavior of the stress field We derive a decompcsition in regular part, edge and vertex singularities.
The spectral problem is solved by a spline - Galerkin method. We present numerical results for vario us geometries, characterizing the leading singular term of the desired stress field.

0. Introduction

Elastic fields governed by the Lame equations will have singular behavior at edges and vertices of the boundary. These singularites pollute the accuracy and convergence of finite difference, finite element and boundary element schemes. For their improvement by grading meshes or augmenting the elements with special singular functions one needs explicit knowledge of the form of the singularities near the edges and vertices (see [4] and [2]). This knowledge is also very important for predicting crack propagation.
We consider the first fundamental problem of elasticity. Boundary integral formulation and a regular transformation of the vicinity of a vertex into a cone yields a fractional convolution as the local form of the equation. Thus, Mellin transformation leads to a holomorphically parameter dependent one - dimensional integral equation over a non smooth curve with the Mellin transformed stress as unknown. The singular expansion terms are determined by the nonlinear eigenvalues and their corrrsponding generalized eigenvectors of the parameter dependent integral equation. For the numerical treatment of the spectral problem we use a spline - Galerkin boundary integral equation method. We show numerical results for circular cones and the well known Fichera vertex.

1. The singular expansion

Let $\Omega \subset R^3$ be a bounded polyhedral domain with boundary Γ. Γ shall be given by the union of piecewise plane boundary pieces Γ_j, $\Gamma = \bigcup_{j=1}^{J} \Gamma_j$. We further assume that Γ may be described in the neighborhood of each vertex by a corresponding conical surface $\gamma \cdot R\!\!I_+$, $\gamma \subset R^2$, where γ is a piecewise circular one - dimensional curve on the unit sphere.
The fist fundamental problem of elasticity is given by

$$\begin{aligned} \mu \Delta \vec{u} + (\lambda + \mu)\, \mathrm{grad}\, \mathrm{div}\, \vec{u} &= 0 \\ \vec{u}_{|\Gamma} &= \vec{\phi} \end{aligned} \tag{1.1}$$

Δ denotes the Laplacian and $\lambda > -\frac{2}{3}\mu$, $\mu > 0$ are the Lamé constants. We assume the displacement $\vec{u}$ has finite energy. $\vec{\phi}$ is a given displacement on the boundary with at least $\vec{\phi} \in H^{\frac{1}{2}}(\Gamma)$. For exterior problems we assume suitable radiation conditions. The solution of (1.1) is determined by the given displacement and the unknown stress on the boundary Γ and can be represented by Betti's representation formula [6] The Somigliana identity

$$\begin{aligned} \int_\Gamma E(y,x)\vec{t}\,(y)\, ds_y = \\ \frac{1}{2}(\varepsilon - 1)\vec{\tau}(x) + \frac{\varepsilon}{2}\vec{\phi}(x) + \int_\Gamma T(y,x)\vec{\phi}(y)\, ds_y \end{aligned} \tag{1.2}$$

yields a boundary integral equation for the unknown stress $\vec{t}$. In short we write

$$A\vec{t} = \vec{f}.$$

The fundamental solution of the Lamé equation is given by

$$E(y,x) = \frac{\lambda + 3\mu}{8\pi\mu(\lambda + 2\mu)} \left\{ \frac{I}{|x-y|} + \frac{\lambda + \mu}{\lambda + 3\mu} \frac{(x-y)(x-y)^T}{|x-y|^3} \right\},$$

and

$$T(y,x) = (T_y[E(y,x)])^T$$

denotes the corresponding boundary traction $\vec{\tau}(x)$ is a rigid body motion. The factor $\varepsilon = 1$ for smooth boundary points must be modified at edges and vertices [5] We remark that the behavior of the solution $\vec{u}$ of (1.1) is connected via Betti's representation formula with the behavior of the stress $\vec{t}$ in (1.2).

Let now χ be a C^∞ function with sufficiently small support V and $\chi = 1$ near the vertex x_0 If we set $\vec{\tau} = \chi\vec{t}$, which is zero outside Ω, in spite of our assumption on Γ equation (1.2) can be written as

$$\int_{R^+}\int_\gamma E(\eta, \frac{r}{s}\xi)\, \vec{s\tau}(s\eta)\, d\!l_\eta \frac{ds}{s} = \vec{h}(r\xi) = \vec{f}(r\xi) - \vec{s}(r\xi). \tag{1.3}$$

with the remainder term $\vec{s}(r\xi) = A(1-\chi)\vec{t}$ Thus, Mellin transformation defined by

$$\hat{g}(\alpha) := \int_{R^+} g(r)\, r^{\alpha - 1}\, dr,$$

applied to the fractional convolution (1.3) yields a holomorphic parameter dependent operator family

$$\begin{aligned} &\hat{A}(\alpha)\hat{\vec{\tau}}(\cdot, \alpha - l) = \hat{\vec{h}}(\cdot, \alpha) \\ &\Longleftrightarrow \\ &\int_\gamma \hat{E}(\eta, \zeta, \alpha)\hat{\vec{\tau}}(\eta, \alpha - l)\, d\!l_\eta = \hat{\vec{h}}(\eta, \alpha), \end{aligned} \tag{1.4}$$

The asymptotic expansion of $\vec{\tau}$ is completely determined by (1.4) [7].

$$\begin{aligned} \vec{\tau} = \sum_{l=1}^{L} \sum_{k=1}^{K_l} \sum_{j=1}^{m_{lk}} c_{lkj}\, r^{-\alpha_l - 1} \cdot \\ \left[\vec{e}_{lkj}\, R_l(\xi) + \sum_{p=1}^{P} \sum_{0 < Im\beta_{pn} < -\frac{1}{2}} d_{pn} \vec{v}_{pn}(\xi) \right] \cdot \\ \log^j r\ \chi(r) + \vec{\tau}_R(r\xi), \end{aligned} \tag{1.5}$$

where r is the distance to the vertex x_0. L is the number of singular functions due to the smooth remainder. α_l is an eigenvalue of $\hat{A}(\alpha)$. The dimension of the geometric eigenspace of $\hat{A}(\alpha_l)$ is denoted by K_l. The maximal length of the chains of generalized eigenfunctions $(\vec{e}_{lk0}, \ldots, \vec{e}_{lk,m_{lk}})$ to an eigenvalue α_l and an eigenfunction $\vec{e}_{lk0}$ is m_{lk}. The number c_{lkj} is a complex constant and $\vec{\tau}_R$ is a smooth remainder.
The expression in brackets belong to the edge singularities which are given by corresponding special two dimensional problems (see [8]). These edge singularities can analytically evaluated and are not of our interest here. The

main problem is to obtain the vertex singular exponent α_0, which can only be done numerically.

2. Numerical approximation of the eigenvalue problem

To approximate $\hat{A}(\alpha)\vec{c} = 0$ we use graded meshes near the corners of γ [1]. The grading exponents are chosen suitable to the singularities of the generalized eigenfunctions, which themselves belong to the edge singularities. Our Galerkin procedure for the approximation of the non linear eigenvalue problem reads as follows [8]

Compute values α^N, such that the linear system

$$\sum_{j=1}^{2NP} c_j \langle \hat{A}(\alpha)\, \mu_j, \mu_i \rangle_0 = 0 \qquad , I = 1, \ldots, 2NP$$

has non trivial solutions $(c_1, \ldots, c_{2NP})$,
where μ_i and μ_j are piecewise constant B - splines on γ.

The value α^N is the approximated eigenvalue of $\hat{A}(\alpha)$. We have proofed the following asymptotic convergence.

$$\left| \alpha_0 - \alpha^N \right| \leq cN^{-\frac{2}{\kappa}},$$

with constant c independent of N, $N \geq N_0$, κ denotes the Riesz - Index of the eigenvalue α_0.

3. Numerical results

First we consider the case when our domain Ω is a circular cone with angle 2Θ:

$$\Omega = \left\{ r \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} \;:\; r \in \mathbb{R},\ 0 \leq \theta \leq \Theta,\ 0 \leq \phi \leq 2\pi \right\}.$$

In the following Figure 1. we show the dependence of the eigenvalues α^N with real part equal zero and smallest absolute value on the angle Θ.

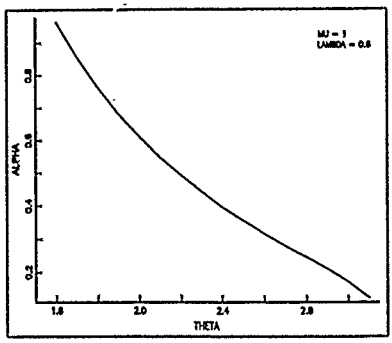


Figure 1. : Eigenvalues of the circular cone

Figure 2. shows the interaction of the the smallest imaginary eigenvalue and the Lamé constants for the well known Fichera vertex [3]

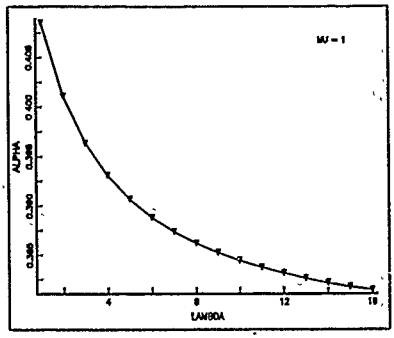


Figure 2. : Eigenvalues of the Fichera vertex

Notice, that the alpha axes in both figures denotes the imaginary part of α

References

[1] G. A. Chandler : : *Galerkin's method for boundary integral equations on polygonal domains* J. Austral. Math. Soc., Ser B. Vol. 8, 1 - 13 (1984).

[2] M. Costabel, E. Stephan : *Boundary integral equations for mixed boundary value problems on polygonal plain domains and Galerkin approximation.* In: Mathematical Models ans Methods in Mechanics 1981, W. Fiszdon and K Wilmánski Eds , Banach Center Publications Vol. 15, 175 - 251 (1985).

[3] G. Fichera : *Numerical and Quantitative Analysis.* Pitman, London (1978).

[4] G. Fix, G. Gulatti, G. I. Wakoff : *On the use of singular functions with finite element approximations.* J. Comp. Phys , Vol. 13, 209 - 238 (1973).

[5] F. Hartmann : *The Somigliana Identity on piecewise smooth surfaces* Journ. of Elast , Vol. 11, 403 - 423 (1981).

[6] G. C. Hsiao, W. L. Wendland : *On a boundary integral method for some exterior problems in elasticity.* Proceedings of Tbilisi University (Trudy Tbiliskogo Univ.) 257, Ser. Mat. Mech Astron , Vol. 18, 31 - 60 (1985).

[7] H. Schmitz : *Über das singuläre Verhalten der Lösungen von Integralgleichungen auf Flächen mit Ecken.* PhD Thesis, Univ. Stuttgart (1989).

[8] K. Volk : *Zur Berechnung von Singulärfunktionen dreidimensionaler elastischer Felder.* PhD. Thesis, Univ. Stuttgart (1989).

EIGENVALUES AND STRESS INTENSITY FACTORS AT REENTRANT EDGES AND CORNERS FOR THREE DIMENSIONAL LAMÉ PROBLEMS

INGO BECKER
Industrieanlagen-Betriebsgesellschaft mbH (IABG)
Einsteinstraße 20
8012 Ottobrunn
Germany

and

ECKART SCHNACK
Institute of Solid Mechanics
Karlsruhe University
Kaiserstraße 12
7500 Karlsruhe
Germany

Abstract

The solution of an elliptic boundary value problem as the Lamé system contains singularities at reentrant edges and corners. The stress intensity factor is given by the coefficient of the known edge singularity. It can be calculated efficiently by a mixed numerical method with special boundary elements around the crack front containing the asymptotic crack solution and with finite elements in the residual domain. The unknown eigenvalue of the corner singularity in case of a surface crack can be determined numerically from the stress field around the crack tip computed by a FE method.

Introduction

Singularities in the solution of elliptic boundary value problems are important for theoretical considerations in mathematics as well as for the calculations in engineering applications. In linear elastic fracture mechanics (LEFM) crack problems are described by the Lamé system for the displacement field u. For three dimensional domains this elliptic boundary value problem contains singularities at reentrant edges and corners. In the interior of a structure cracks are forming reentrant edges with an opening angle 2π. At the intersection of a crack with a free surface a reentrant corner exists. At the intersection point a corner singularity is superposed the edge singularity of the intersecting crack front.

Edge singularity

For cracks in the interior of a structure the asymptotic solution is well known. The stress field near the crack tip behaves like $1/\sqrt{r}$. Important parameters to calculate for fracture mechanics are the stress intensity factors given by the coefficients of the $1/\sqrt{r}$ singularity. The numerical calculation of the stress intensity factors in this case can be done very efficiently by a mixed method using trial functions fulfilling the equilibrium and Neumann boundary condition [1], [2] [3] realized by the eigenfunctions of the solution for the semi-infinite crack [4].

The considered domain is devided into a region around the crack front Ω_1 with elastic fields fulfilling the equilibrium and the Neumann condition exactly and a residual domain Ω_2 where a finite element formulation is used. Applying the principal of virtual work a generalized equilibrium condition is obtained.

$$\int_{\Omega_2} \sigma\, \varepsilon^*\, d\Omega - \int_{\Gamma_{N_2}} \Gamma u^*\, d\Gamma + \int_{\Gamma_1} t\, u^*\, d\Gamma - \int_{\Gamma_{N_1}} \Gamma u^*\, d\Gamma = 0$$

where σ and ε denote the stress and strain tensor, t and u the traction and displacement field and Γ^* the given traction as Neumann condition. The test functions are marked by a star. The surface Γ_N represents that part of the boundary with the Neumann condition. In addition there results a generalized continuity condition:

$$\int_{\Gamma_1} (u_\Omega - u_\Gamma)\, t^*\, d\Gamma = 0$$

where u_Ω denotes the displacement field fulfilling the equilibrium and Neumann condition and u_Γ an additional displacement field on the boundary of the crack region, which is compatible to the adjacent finite elements.

For the discretization of the crack region trial functions $u_\Omega = L\beta$ and $t = R\beta$ are formulated on the surface of macro elements with free parameters β. The matrices L and R contain functions fulfilling the equilibrium and Neumann condition realized by the eigenfunctions of the solution for the semi-infinite crack. Additionally the displacement field $u_\Gamma = Nd$ is formulated with the interpolation matrix N and the nodal displacements d. As result a stiffness relation with the following stiffness matrix for a macro element in the crack region is obtained:

$$k_1 = T^T H^{-1} T$$

with

$$T = \int_{\Gamma_1} R^T N\, d\Gamma \quad \text{and} \quad H = \int_{\Gamma_1} R^T L\, d\Gamma$$

which is symmetric and positiv definite. Therefore these macro elements can be easily combined with other finite elements. This method has additionally the advantage of a boundary element method because for the macro elements only surface integrals has to be calculated.

As a test example an elliptical crack under tension was calculated with a ratio of the main axis of 3. Because of symmetry only 1/8 of the domain has to be considered. The region around the crack front was discretized by 6 pairs of hexahedral macro elements. For the residual domain about 300 tetrahedral elements with quadratic displacements were used. The numerical result for the stress intensity factor along the crack front (Fig. 1) agrees well with the analytical values for an infinite body. The maximum deviation is aout 6%.

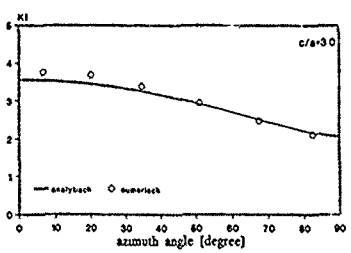


Fig. 1: Stress intensity factor along the crack front

Corner singularity

The intersection point of a crack front with a free surface forms a reentrant corner with the crack front as an incoming reentrant edge. At this point an unknown corner singularity is superposed to the known edge singularity. In this case the domain can be described by a polyhedron in $\mathbb{R}^3$. The following decomposition in a polyhedral cone with the spherical coordinates (r,θ_1,φ_1) has been proved by M. Dauge [5] and von Petersdorff [6]:

$$u = u_0 + \chi(r) \sum_{\lambda_k} a_k r^{\lambda_k+1} v_k(\theta_1,\varphi_1)$$
$$+ \sum_{j=1}^{J} \chi_j(\theta_j) \sum e^j_{mp}(r)\, \theta_j^{2p}\, g^j_m(\theta_j,\varphi_j)$$
$$+ \left[\chi(r) \sum_{\ell=1}^{L(\lambda)} \sum_{q=0}^{\ell} r^{\lambda_k+1} \log^q r\, \hat{v}_{k\ell q}(\theta_1,\varphi_1) \right]$$

The corner singularities are given by the eigenvalues λ_k while the sum over j describes the edge singularities. In special cases there appear logarithmic terms. The functions χ are cut-off functions and u_0 represents the regular part. For the surface crack this leads near the free surface to the following expression:

$$u = u_0 + \sum_{\lambda_k} a_k r^{\lambda_k+1} \omega_k(\theta,\varphi) + \rho^{1/2}\Big(\sum_{\lambda_k} a_k r^{\lambda_k+1/2} + c(z)\Big) f(\varphi)$$

with $c(z) \longrightarrow 0$ for $z \longrightarrow 0$, where the z-axis is given by the crack front and ρ denotes the distance to it. The displacement field u can be decomposed into a corner singularity given by λ_k and an edge singularity of $\sqrt{\rho}$-type, where the coefficient can be interpreted as a stress intensity factor. This stress intensity factor depends on λ_k and tends to zero for $z \longrightarrow 0$ when the smallest $\lambda_k > -\frac{1}{2}$. Otherwise, it increases to infinity approaching the free surface.

The unknown eigenvalue λ_k of the corner singularity can be determined numerically from the stress field around the crack tip computed by FEM. If logarithmic singularities are absent the asymptotic solution for the stresses has the following form:

$$\sigma_{ij} = r^{\lambda}\, f_{ij}(\theta,\varphi).$$

Plotting the stresses versus the distance for constant polar angles θ and φ in a double logarithmic scale one gets a straight line with the unknown eigenvalue as the slope. In Fig. 2 the discretization of 1/8 of a structure with a crack front coming perpendicular to a free surface is shown. The discretization into 2300 tetrahedrons results in about 11000

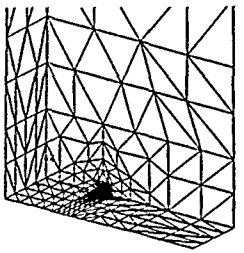


Fig. 2: Discretization for a surface crack

equations. The stresses on the free surface Fig. 3 are plotted as a function of the distance from the crack tip along two lines of $\varphi = 45°$ and $\varphi = 60°$ in a logarithmic scale (Fig. 4). From the slope of the lines one obtains the eigenvalue of the corner singularity:

$$\lambda = -0.40 \pm 0.02$$

This value is clearly above $-\frac{1}{2}$, which implies a vanishing stress intensity factor at the free surface. A possible logarithmic singularity would result in a curvature in the plot, which cannot be observed in this example.

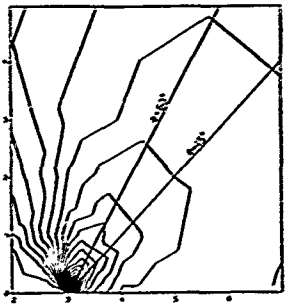


Fig. 3: Isolines of the stresses on the free surface

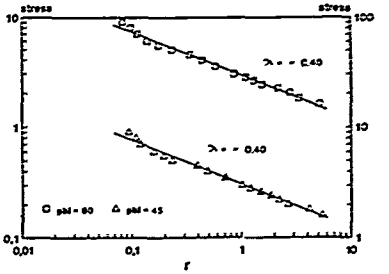


Fig. 4: Plot of the stresses as a function of r

References

[1] Schnack, E., Wolf, M.: Application of displacement and hybrid-stress methods to plane notch and crack problems. Int. J. Num. Meth. Engng. 12, 963-975 (1978).

[2] Becker, L., Karaomanoglu, N., Schnack, E.: Mixed methods with BEM for three dimensional fracture mechanics. In Brebbia, C. A., Wendland, W. L., Kuhn, G. (eds.): Boundary elements IX 2, 227-241 (1987).

[3] Becker, L.: Numerische Berechnung von Ecken- und Kantensingularitäten elastischer Felder für dreidimensionale Rißprobleme. Dissertation Universität Karlsruhe (1989).

[4] Hartranft, R. J., Sih, G. C.: The use of eigenfunction expansions in the general solution of three-dimensional crack problems. J. Math. Mech. 19, 123-138 (1969).

[5] Dauge, M.: Régularités et singularités des solutions de problèmes aux limites elliptiques sur des domaines singuliers de type à coins. Thèse Univ. de Nantes (1986).

[6] v. Petersdorff, T.: Randwertproblem der Elastizitätstheorie für Polyeder - Singularitäten und Approximation mit Randelementmethoden. Dissertation TH Darmstadt (1989).

COMPUTERIZED RAY TRACING MODEL TO AID ELECTROSTATIC LENS DESIGNING

L. Kiss
Physics Department of Jozsef Gyula College
Boldogasszony sgt. 6., PO Box 396, H-6701 Hungary

Abstract - A computerized ray tracing model for systematic investigation of axially symmetric electrostatic lenses from the point of view of the spherical and chromatic aberrations is proposed. The calculation of the first order optical properties of the lenses and their spherical and chromatic aberrations is based on the direct numerical solution of the trajectory equation for paraxial, nonparaxial and chromatic rays. The axial potential distribution of the focusing field is constructed as a fifth order spline. The two-interval fifth order spline lenses have been studied in some detail and the experiences of the application of the model are presented in the paper.

I. INTRODUCTION

Axially symmetric electrostatic lenses are very important in the instruments of the different electron and ion beam technics serving the purposes of surface science, analytical chemistry, electron and ion beam lithography or ion implantation, for instance. Consequently, the developments and the utilizations of new computational methods and computer applications for designing prime electrostatic lenses with rotational symmetry are essential. Since the spherical and chromatic aberrations are the main limitations of the optical quality of the electrostatic lenses, we select the spherical and chromatic aberration coefficients in infinite magnification mode, referred to the object space and related to the object side focal length as figures of merit to describe the lens performance. Currently, beside the advanced computerized techniques of lens analysis [1] - [3], a very effective approach to design axially symmetric electrostatic lenses which provide usable optical properties in first order and also the best figures of merit is electrostatic field optimization and synthesis of lenses [4], [5]. The essence of ion optical optimization and synthesis is that any imaging electrostatic field, its optical properties, aberrations and the applied figures of merit, together with the electrode configurations of the lens making possible the imaging field are totally determined by the axial potential distribution of the field. Consequently, instead of analyzing a large number of different electrode and pole piece configurations varying the geometrical parameters and the electrode voltages to find lenses with higher performance, it is advantageous to investigate axial potential functions to select the optimum axial distribution. Obviously, in the approach of synthesis it is essential how the axial potential distribution is defined and built up. The application of cubic splines has yielded a very effective possibility of the utilization of ion optical synthesis for practical lens optimization and design [6] - [9]. In this paper we construct the axial potential distribution as a fifth order spline to improve our ray tracing model [10] originally based on the cubic spline technique.

II. RAY TRACING MODEL

In case of electrostatic lenses, the meridional motions of charged particles can be considered to calculate the first order optical properties and the spherical and chromatic aberrations without violating the generality of the problem. For meridional motions, the following universal differential equation [10] determines the actual particle trajectories:

$$r'' = \frac{1 + (Q/m_0)(u - u_p)/c^2}{[2 + (Q/m_0)(u - u_p)/c^2](u - u_p)} \cdot (1 + r'^2)\left[\frac{\partial u}{\partial r} - r'\frac{\partial u}{\partial z}\right]. \qquad (1)$$

where Q is the particle's charge, m_0 is the particle's rest mass, $u = u(r,z)$ is the potential of the focusing field, $r = r(z)$ is the particle trajectory, u_p is the potential where the velocity of the particle is zero, and c is the speed of light in vacuum. The cylindrical coordinate system is used and the primes represent differentiation with respect to z which is the independent variable along the optical or symmetry axis of the lens. The potential can be expressed as a power series in r in the form of [11]

$$u(r,z) = U(z) - (r^2/4)U''(z) + (r^4/64)U''''(z) - \cdots, \qquad (2)$$

where $U = U(z)$ is the axial potential distribution of the field. If the axial potential U built up as a cubic spline the third term in Eq. (2) will be a delta function. This fact can yield some inaccuracy for the third order spherical aberration (the error of the calculation can be estimated by using integral expressions). When a fifth order spline model is utilized the forth order terms are continuous, therefore inaccuracies can occur only in higher order approximation. The differential equqtions for paraxial, nonparaxial, and chromatic rays can be given by only substituting the power series of $u(r,z)$ in the appropriate order into Eq. (1).

III. RESULTS

A systematic analysis of the unipotential

and retarding two-interval spline lenses have been carried out by using the model outlined in Sec. II. In case of unipotential lenses, the spherical figure of merit C_{so}/f_o as the function of the axial potential at the middle point of the lens U_{mid} is shown by solid curve in Fig. 1. U_{mid} is related to the value of the axial potential at the object side lens endpoint. The dashed curve shows the spherical figure of merit calculated by utilizing the ray tracing model based on cubic splines [12].

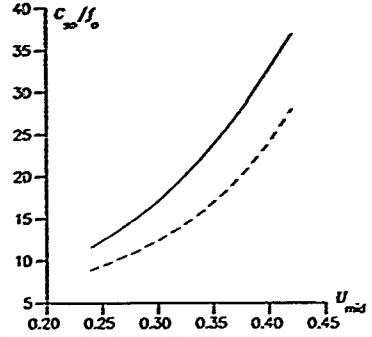


Fig. 1. C_{so}/f_o as the function of the axial potential at the lens midpoint U_{mid}. The solid and the dashed lines are for the fifth and the third order spline models, respectively.

In a special case of retarding lenses, when an approximately nine times change in the kinetic energy of the particles is considered, the spherical figure of merit is shown in Fig. 2.

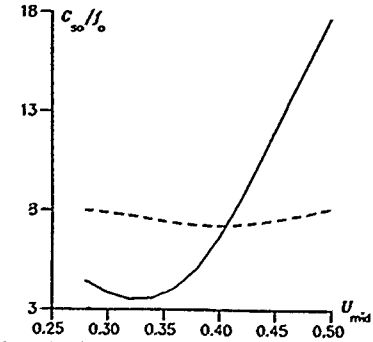


Fig. 2. C_{so}/f_o as the function of the axial potential at the lens midpoint U_{mid}. The solid and the dashed lines are for the fifth and the third order spline models, respectively.

The results seem to indicate that the spherical figure of merit has a similar trend in cases of the fifth order and the cubic splines. However, the corresponding particular values are significantly different. In case of the retarding lenses the minimum of the curve of the spherical figure of merit determined by using the fifth order spline model is much more expressive than which is calculated by using cubic splines.

ACKNOWLEDGMENTS

The Individual Mobility Grant from TEMPUS, making possible the presentation of the paper at the 13th IMACS World Congress in Dublin, is highly appreciated and gratefully thanked by the author.

REFERENCES

[1] E. Munro, *Finite difference programs for computing tolerances for electrostatic lenses*, J. Vac. Sci. Technol., vol. B6, p. 941, 1988.
[2] T. Tsumagari, H. Chiwa, and T. Noda, *Design of a low-aberration lens for focused ion beams*, J. Vac. Sci. Technol., vol. B6, p. 949, 1988.
[3] E. Munro, J. Orloff R. Rutherford, and J. Wallmark, *High-resolution, low-energy beams by means of mirror optics*, J. Vac. Sci. Technol., vol. B6, p. 1971, 1988.
[4] M. Szilagyi, *Synthesis of Electron Lenses*, in Electron Optical Systems for Microscopy, Microanalysis, and Microlithography, Scanning Electron Microscopy Inc., AMF O'Hare Il., 1984, p. 75.
[5] M. Szilagyi, *Electron Optical Synthesis and Optimization*, Proc. IEEE, vol. 7, p. 412, 1985.
[6] M. Szilagyi, *Electrostatic spline lenses*, J. Vac. Sci. Technol., vol. A5, p. 273, 1987.
[7] M. Szilagyi and J. Szép, *A Systematic Analysis of Symmetric Three-Electrode Electrostatic Lenses*, IEEE Trans. Electron Devices, vol. ED-35, p. 2834, 1987.
[8] L. Kiss, *Computerized investigation of electrostatic lens potential distributions*, in Proc. 12th IMACS World Congress, Paris, Gerfidn, France 1988, p. 242.
[9] J. P. Adriaanse, H. W. G. van der Steen, and J. E. Barth, *Practical optimization of electrostatic lenses*, J. Vac. Sci. Technol., vol. B7, p. 851, 1989.
[10] L. Kiss, *Electrostatic lens potentials with small relativistic spherical aberration*, Rev. Sci. Instrum., vol. 60, p. 907, 1989.
[11] A. B. El-Kareh and J. C. J. El-Kareh, *Electron Beams Lenses and Optics*, vol. 1, Academic Press, New York, 1970, p. 14.
[12] L. Kiss and J. Krafcsik, *Optimization and synthesis of electrostatic non-accelerating two-interval spline lenses*, J. Vac. Sci. Technol. B, March, 1991.

POWER FLOW MODELLING OF COMMERCIAL VEHICLE AIR BRAKE SYSTEMS

V.MOHAN
Member R & D
Sundaram Clayton Ltd.
Madras INDIA

Dr.M.SINGAPERUMAL
Asst. Professor
Dept. of Mechanical Engg.
IIT, Madras INDIA

ABSTRACT

Power flow approach has been chosen for modelling the dynamic response of a commercial vehicle air brake system. Most of the published work deal with systems where compressibility effects are neglected. In this paper simpler constitutive relationships are adopted for resistance and capacitance fields having pressure, mass flow rate and temperature as parameters. With these elements, it is possible to represent actual pneumatic systems as lumped model systems. The control valve and the pipe are modelled as equivalent orifices. The orifice parameters namely, the sonic conductance 'c' and the critical pressure ratio 'b' are estimated by a simple transient technique. The brake response in terms of pressure rise in the brake chamber is studied. Actual field experiments are conducted to measure the dynamic response. A good correlation has been obtained between the experimental results and the simulated model.

1.0 INTRODUCTION

Dynamic response is very important in air brake actuation and has a direct bearing on the stopping distance. Ultimate safety of the vehicle depends upon the stopping distance. There are also certain mandatory regulations to be complied with, by the braking systems. There is hardly any work reported on automotive air brake systems response studies. These response characteristics can be best obtained through mathematical models.

Of the many methods, the power bond graph is chosen because of its superiority in representing physical action dictating the dynamic response by means of compact symbolic notation and ease of modification. Further, the brake system actuation is characterised by dynamic energy flows. Hence bondgraph representation is very relevant in this context. While it is well established for the oil hydraulic systems, the use of bond graph for the pneumatic system is still under development. So an attempt is made here to develop a bond graph formulation of an air brake system.

Air brake systems use compressed air to apply the brakes. Air is compressed by a compressor driven by the engine and stored in a reservoir. When the driver depresses the foot pedal, air from the reservoir flows to the brake chambers at a pressure determined by driver's pedal effort and displaces the diaphragms in the chambers. The movement of the diaphragms is transmitted to the brake shoes, forcing them against the drums and thus braking the wheels.

The dynamic response of a complete brake system consists of a quasi-static component and a transient component. The transient behaviour is associated with rapidly changing system variables, such as brake line pressure following a rapid pedal force input. The quasi-static behaviour is associated with slowly changing variables, such as the change in coefficient in friction between lining and rotor due to change in wheel speed during the deceleration of the vehicle. In this only the transient behaviour is modelled. The actual stopping distance is affected by time delays required to apply brakes and to build up the brake force. Generally the actuation time, is small compared to the build up time and hence the build up time is alone modelled.

2.0 MODELLING OF THE FLOW PROCESS

The flow through the brake pipes and valves can be modelled either as a continuous model or as a lumped model. During the initial design of layouts, many configurations require quick evaluation. So a lumped model approach is favoured as continuous models are highly time consuming. To arrive at a suitable lumped system, Power bond approach has been adopted.

The pneumatic medium is characterised by significant changes in pressure, density and temperature.

KARNOPP [1] has shown that for a compressible fluid flow process it is necessary to use the stagnation enthalpy (h) as the effort variable and the mass flow rate ($\dot{m}$) as the flow variable to properly account for real flow. CHENG [2] modified the above equation and developed a set of multiport elements for pneumatic systems. The enthalpy (h) is expressed in terms of pressure (p) and the temperature (T) and they are used as effort variables. This approach with suitable modifications has been adopted for the brake system simulation. A simplified layout of an air brake system is shown in Fig. 1. When the valve is operated, air flows from the reservoir to actuator. The volume of the actuator is nearly constant. Resistance is offered by the valve and also by the pipe. When the valve is fully opened, the resistance offered by the valve and the pipe can be considered constant.

The steps involved are:

-To find suitable method to specify valve as an equivalent orifice
-To find equivalent orifice for pipes
-To develop power flow model of the system and numerically integrate the equations derived from the model

2.1 THE SYSTEM BOND GRAPH

The lumped model representation of the simplified air brake system is shown in Fig. 2. The brake application valve is represented as an orifice whose conductance 'c' and critical pressure ratio 'b' is estimated through a transient technique [3]. The pipes are also represented by equivalent orifices whose c and b values are obtained using an empirical relationship [4]. The pipe volumes are lumped with actuator volume as capacitance. The system equations derived from bond graph are listed below with the parameter values.

Notation

C_1-conductance of brake application valve in $dm^3/s.bar$
c_x-conductance of pipes connecting to front brake chamber in $dm^3/s.bar$
c_y-conductance of pipes connecting to rear brake chamber in $dm^3/s.bar$
p_o-reservoir pressure in bar
p_1-pressure at pipe junction
p_2-pressure in front brake chamber in bar
p_3-pressure in rear brake chamber in bar
m_x-mass in kg
$\dot{m}_o$-mass flow rate from reservoir to pipe junction in kg/s
$\dot{m}_x$-mass flow rate from junction to front brake chamber in kg/s
$\dot{m}_y$-mass flow rate from junction to rear brake chamber in kg/s
$\dot{m}_1$-net mass flow rate from reservoir to pipe junction in kg/s
T-temperature in k
C_x-capacitance
R -gas constant =287 Nm/kg k
V_x-volume in m^3
t -time in seconds
ρ_o - Density of air kg/ m^3; 20 C= 1.28 kg/m^3
n -polytropic constant
T_o-Temperature 20 C

Equations

$$\dot{m}_2 = c_1 p_o \sqrt{\frac{T_o}{T_1}}$$

$$\dot{m}_2 = c_1 p_o \sqrt{\frac{T_o}{T_1}} \sqrt{1 - \left(\frac{\frac{p_1}{p_o} - b_1}{1 - b_1}\right)^2} \quad (1) \quad \left(\text{when } \frac{p_1}{p_o} > b_1\right)$$

$$\dot{m}_y = C_y P_y \sqrt{\frac{1}{T_y}}$$

$$\dot{m}_y = C_y P_y \sqrt{\frac{1}{T_y}\left[1-\left(\frac{\frac{P_y}{P_y}-b_y}{1-b_y}\right)^2\right]} \quad (2) \left(\text{when } \frac{P_y}{P_y} > b_y\right)$$

$$\dot{m}_y = C_y P_y \sqrt{\frac{1}{T_y}}$$

$$\dot{m}_y = C_y P_y \sqrt{\frac{1}{T_y}\left[1-\left(\frac{\frac{P_y}{P_y}-b_y}{1-b_y}\right)^2\right]} \quad (3) \left(\text{when } \frac{P_y}{P_y} > b_y\right)$$

$$\dot{m}_i = \dot{m}_o - \dot{m}_y - \dot{m}_y \quad (4)$$

$$P_i = P_i(0) + \int \frac{1}{C_{fl}} \dot{m} \, dt \quad ; C_{fl} = \frac{V_i}{RT_i} \quad (5)$$

$$T_i = T_i(0)\left(\frac{P_i}{P_i(0)}\right)^{(n-1)/n} \quad (6)$$

Initial values
Pressure in gauge values
$p_s(0) = 7.5$ bar
$p_1(0) = 0$ bar
$p_2(0) = 0$ bar
$p_3(0) = 0$ bar
T (0) = 300 K
PARAMETER VALUES
c_1=2.425 (while charging)
= 3.19 (while discharging)
b_1=0
c_2=5.8567
b_2=0.2739
c_3=3.010
b_3=0.3217
n=1.4
$V_2 = 0.723 * 10^{-3}$ m³
$V_3 = 1.359 * 10^{-3}$ m³
$V_1 = 0.423 * 10^{-3}$ m³
These equations are numerically integrated to obtain the system response.

The theoretical response are compared with the experimental results obtained from a vehicle brake. (Fig. 3 and 4)

2.2 RESULTS AND DISCUSSION

The results for the front and rear brake chambers matches well with simulated values especially while charging. The time taken to reach 75 % of the asymptotic pressure is 0.3 seconds and the corresponding simulated value is also 0.3 seconds. The correlation is better in the front brake chamber. To improve the dynamic model, the flow to the rear portion which is governed by the pressure at the junction may be modelled with some weighting function to account for the some back flow that may take place from the front portion.

3.0 CONCLUSIONS

The possibility of using power bond graph to obtain the lumped model of commercial vehicle air brake systems is presented. Relevant equations and the way to specify the parameters are discussed. The theoretical results compare well with experimental data. This work could form a basis for extending the model to include the actuating valve dynamics and also to the frequency domain analysis and sensitivity analysis of brake systems with a view of improving their dynamics, which are critical in the new antilock brake systems.

REFERENCES

1.Karnopp,Dean (1972) - "Bond Graph for Fluid Dynamic Systems",Journal of Dynamic Systems,Measurement and Control Transactions of ASME, Sep.

2.Cheng,R.M.H and Harris,O.G (1986)- "A power bond graph approach for modeling pneumatic control systems",- Journal of Fluid Control, vol.16,no.4

3.Mohan, V.,(1990) - "Dynamic analysis of commercial vehicle air brake systems" -M.S. Dissertation ,Indian Institute of Technology, Madras.

4.Bata ,H.P. (1985) "Flow measurements and technical flow data" ,o+p oilhydraulik und pneumatik, 29 (1985),no. 7 p 541

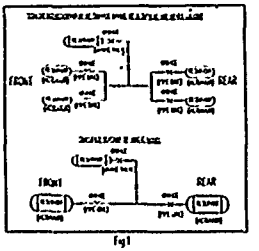


Fig1

BOND GRAPH REPRESENTATION OF THE LUMPED MODEL

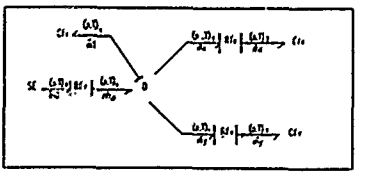


Fig 2

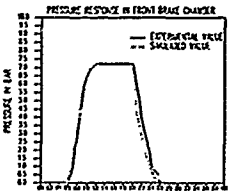


Fig 3

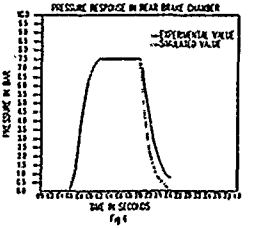


Fig 4

A CONTRIBUTION TO THE NUMERICAL STUDY OF THE VAPOR FLOW CHARACTERISTICS OF SLENDER CYLINDRICAL HEAT PIPES

by

Basem S. Attili
K.F.U.P.M.
P.O.Box 1927
Dhahran 31261, Saudi Arabia

Abstract: The Navier- Stocks equation and the continuity equations which are applicable to the laminar vapor flow in slender cylindrical heat pipes are solved numerically. An implicit finite difference method is used to solve the equations. The pressure gradient term is eliminated, which reduces the number of unknowns by one. The effect of pressure drops is then studied for different Reynolds numbers.

1 INTRODUCTION

Through a porous wall, the vapor flow dynamics in the evaporator and condenser of a heat pipe is very much similar to pipe flow with injection and suction. The principle of thermal fluid governs the performance of the heat pipe. The reason being that the heat pipe is a heat transfer device. Good attention was given to the fluid mechanical aspects of heat pipes. See for example the review given by Tien[2].

To obtain steady state operation of a heat pipe, the sum of the axial pressure difference due to body forces, the pressure along the flow line due to surface tension and both the liquid flow and the vapor flow gradients must all add up to zero. It should be noted that relativly speaking, the fluid dynamics of vapor flow is not so easy even if the geometry and boundary specification of the problem are simple.

Based on the radial flow of the vapor at the wall of a heat pipe, a radial Reynolds number Re is introduced. It is a well known fact that if the absolute value of Re is much less than one, viscous effects dominates, while if Re is much bigger than one, momentum forces dominate.

Theoretical invistigations have been done to steady incompressible laminar flow. White[4] have found similarity solutions but Masuzawa, Tanahashi and Ando[5], Gupta[6] and Hornbeck, Rouleau and Osterle[7] found entry region solutions. To summarise the results for different Re. There will be dual solutions at the walls for Re less than 2 3 and for values of Re between 9.1 and 20 6 . Multiple solutions are obtained when Re is 20.6 and one solution occurs when Re is 2.3 . The flow approaches poiseuille form when Re approaches zero.

2 BASIC EQUATIONS AND ASSUMPTIONS

We will assume steady, incompressible, axisymmetric and laminar flow. The pressure distribution depends on axial coordinates. Also, we will assume constant radial velocity through the pipe wall and the radial velocity component compared to the average axial velocity is small.

The governing equations are.

Axial momentum equation

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial r}=\frac{-1}{\rho}\frac{dp}{dx}+\nu\left(\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}\right). \tag{2.1}$$

Continuity equation

$$\frac{\partial u}{\partial x}+\frac{1}{r}\frac{\partial(vr)}{\partial r}=0. \tag{2.2}$$

Subject to the boundary conditions,

exit wall conditions

$$i)\quad u(0,r)=u(1,r)=0,$$

$$ii)\quad v(0,r)=v(1,r)=0,$$

axial conditions

$$iii)\quad v(x,0)=0,$$

$$iv)\quad \frac{\partial u}{\partial r}(x,0)=0,$$

no slip condition

$$v)\quad u(x,r_0)=0,$$

and

$$vi)\quad v(x,r_0)=c \quad \text{(constant)} \tag{2.3}$$

where

$$\begin{cases} c>0 & \text{Injection.}\\ c=0 & \text{Adiabatic zone.}\\ c<0 & \text{Suction.}\end{cases}$$

This approximation is made for positions of x which are more than one radius from the end. This is since the pressure gradient term may not be negligible. See Busse[1]. Equations (2.1) and (2.2) are two equations in three unknowns, namely u, v and p. To solve numerically, one must eliminate the pressure term. To do so, multiply equation (2.1) by r and integrate with respect to r from $r=0$ to $r=r_0$ to obtain

$$-\frac{1}{\rho}\frac{dp}{dx}=\frac{2}{r_0^2}\frac{\partial}{\partial x}\int_0^{r_0}u^2 r\,dr-\frac{2\nu}{r_0}\frac{\partial u}{\partial r}\Big|^{r_0}, \tag{2.4}$$

which when substituted in (2.1) leads to

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial r}=\frac{2}{r_0^2}\frac{\partial}{\partial x}\int_0^{r_0}ru^2dr-\frac{2\nu}{r_0}\frac{\partial u}{\partial r}\Big|^{r_0}+\nu\left(\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}\right). \tag{2.5}$$

The details of the above procedure will be presented elsewhere.

Using the dimensionless parameters

$$\bar{u}=\frac{u}{u_a},\bar{v}=\frac{vr_0}{\nu},\bar{r}=\frac{r}{r_0}\quad\text{and}\quad\bar{x}=\frac{x\nu}{r_0^2u_a}, \tag{2.6}$$

equations (2.2),(2.3) and (2.5) can be written in the dimensionless form

$$\bar{u}\frac{\partial\bar{u}}{\partial\bar{x}}+\bar{v}\frac{\partial\bar{u}}{\partial\bar{r}}=-2\frac{\partial\bar{u}}{\partial\bar{r}}\Big|_1+4\int_0^1\bar{r}\bar{u}\frac{\partial\bar{u}}{\partial\bar{x}}d\bar{r}+\left(\frac{\partial^2\bar{u}}{\partial\bar{r}^2}+\frac{1}{\bar{r}}\frac{\partial\bar{u}}{\partial\bar{r}}\right), \tag{2.7}$$

$$\bar{r}\frac{\partial\bar{u}}{\partial\bar{x}}+\frac{\partial(\bar{v}\bar{r})}{\partial\bar{r}}=0, \tag{2.8}$$

subject to the boundary conditions,

$$i)\quad \bar{u}(0,\bar{r})=\bar{v}(0,\bar{r})=0,$$

$$ii)\quad \bar{v}(s,\bar{r}) = \bar{v}(s,\bar{r}) = 0,$$

$$iii)\quad \bar{v}(\bar{z},0) = 0,$$

$$iv)\quad \frac{\partial \bar{u}}{\partial \bar{r}}(\bar{z},0) = 0,$$

$$v)\quad \bar{v}(\bar{r},1) = 0,$$

and

$$vi)\quad \bar{v}(\bar{z},1) = \begin{cases} -Re & \text{Evaporator.} \\ 0 & \text{Adiabatic zone.} \\ +Re & \text{Condenser.} \end{cases} \qquad (2.3)$$

3 NUMERICAL SOLUTION, RESULTS AND CONCLUSIONS

The numerical procedure which will be used is an indirect modification of the marching procedure used by Hornbeck, Rouleau and Osterle[7]. It is an implicit finite difference scheme. A mesh point is of the form $(\bar{z},\bar{r}) = (mh, nk)$, where h and k are the step sizes in the $\bar{z},\bar{r}$ directions respectively. The scheme is obtained using the following representations

$$\frac{\partial \bar{u}}{\partial \bar{r}} = \frac{\bar{u}^n_{m+1} - \bar{u}^n_m}{h},$$

$$\frac{\partial \bar{u}}{\partial \bar{r}} = \frac{\bar{u}^{n+1}_{m+1} + \bar{u}^{n+1}_m - \bar{u}^{n-1}_m - \bar{u}^{n-1}_{m+1}}{4k},$$

$$\frac{\partial \bar{v}\bar{r}}{\partial \bar{r}} = \frac{\bar{v}^{n+1}_{m+1}\bar{r}_{m+1} - \bar{v}^n_{m+1}\bar{r}_m}{k},$$

and

$$\frac{\partial^2 \bar{u}}{\partial \bar{r}^2} = \frac{\bar{u}^{n+1}_{m+1} + \bar{u}^{n+1}_m - 2\bar{u}^n_{m+1} - 2\bar{u}^n_m + \bar{u}^{n-1}_{m+1} + \bar{u}^{n-1}_m}{2k^2}. \qquad (3.1)$$

For the approximation of $\partial \bar{u}/\partial \bar{r}$ at $\bar{r} = 1$, we used a five point backward formula since the rate of change of the velocity gradient at the wall is large.

The system of difference equations was solved using Gaussian-elimination with full pivoting to obtain $\bar{u}^n_{m+1}$, which is used in the continuity equation to obtain $\bar{v}^n_{m+1}$. The pressure drop gradient is then evaluated from the dimensionless form of (2.4).

The results obtained agrees very much with Hornbeck[8] and Sparrow[3]. For uniform and parabolic profiles, it was noticed that for the case of injection ($Re = -5$), the fully developed normalized velocity at the center ($R = 0$) is between the two profiles, while in the case of suction ($Re = 1.$), it was higher than the two profiles, see Figures 1 and 2. The pressure drop for injection ($Re = -1$) and suction ($Re = 1.$) was analyzed. It was noticed that both the frictional and static pressure drop decreases along the pipe length continously, while the momentum pressure drop decreases for the injection case and increases for the suction case, see Figures 3 and 4. Comparison of the two cases of suction and injection also shows that the pressure drop rate for injection is more than that of suction. This is due to the increase in axial velocity in the injection case.

The velocity gradient was studied for different values of Re. No solution was obtained for $Re > 2.3$ as was observed by White[4]. At the wall, the garadient approaches zero and the separation occurs early, see Figure 5.

More studies are being done on the velocity gradient, vapor pressure drop in asymmetric heat pipe and on a heat pipe with adiabatic zone. The results will be presented elsewhere.

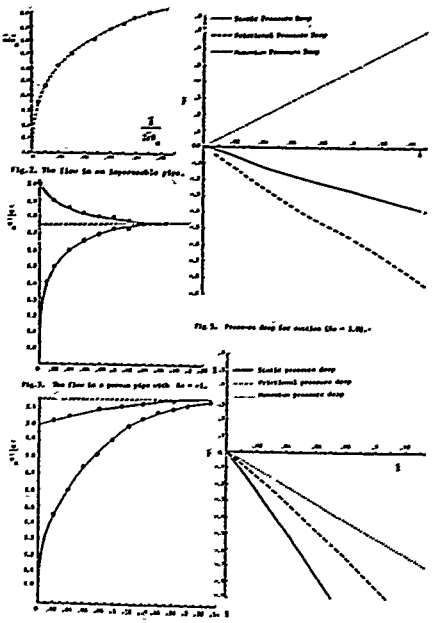


Fig.2. The flow in an impermeable pipe.

Fig.3. Pressure drop for suction ($Re = 1.0$).

Fig.4. The flow in a porous pipe with $Re = 1.0$ (Suction)

Fig.5. Pressure drop for injection ($Re = -1$)

AKNOWLEDGMENT The author would like to aknowledge K.F.U.P.M. for the support he has while doing this research and for using the facilities of the Mathematics department.

References

[1] C. A. Busse, *Pressure Drop in the Vapor Phase of Long Heat Pipes*, Proceedings of the Thermionic Conversion Specialist Conference, Palo Alto, Calif.(1967),381-389.

[2] C.L. Tien, *Fluid Mechanics of Heat Pipes*, Annual Review of Fluid Mechanics, 7(1975), 167-174.

[3] E. M. Sparrow, S. h. Lin, *Flow Development in the Hydrodynamic Entrance Region of Tubes and Ducts*, Phys. Fluids, 7(1964), 338-347.

[4] F. M. White, *Laminar Flow in a Uniformly Porous Tube*, J. Appl. Mech , 29(1962), 201-204.

[5] J. Masuzawa, T. Tanahashi and T. Ando, *Flow of the Entrance Region in a Porous Pipe*, Bull. JSME, 23(1980), 672-678.

[6] R. C. Gupta, *Laminar Flow in the Inlet Region of a Porous Tube*, Appl. Sci. Res., 22(1970), 360-365.

[7] R. W. Hornbeck, W. T. Rouleau and F. Osterle, *Laminar Entry Problem in Porous Tubes*, Phys. Fluids, 6(1980), 672-678.

[8] R. W. Hornbeck, *Laminar Flow in the Entrance Region of a Pipe*, Appl. Sci. Res. Sect., 13(1964), 224-232.

BALANCED MODELS BUILDING WITH GUARANTEED STRUCTURAL PARAMETERS: A SIMULATED ANNEALING APPROACH

L. Fortuna, G. Nunnari, S. Baglio, S. Graziani

Dipartimento Elettrico, Elettronico e Sistemistico
University of Catania

Viale A. Doria 6, I-95125 , Catania, ITALY

Abstract - In this paper an optimization procedure, based on the simulated annealing strategy, is proposed in order to assure a suitable solution to the problem of built-in systems with assigned invariant parameters. The proposed approach is very attractive due to its capability to avoid local minima.

I - INTRODUCTION

The possibility of assigning, in open loop scheme, for SISO systems, both the eigenvalues and other invariant quantities, characterizing some structural properties, such as the stability, controllability and observability, has been previously studied by the authors [1]. The same problem has been also investigated for MIMO systems [2]; in this second case more parameters can be freely chosen making the problem less hard. In any case the existence of the problem solution cannot be proved.
It was proved [1] that, for SISO systems, if an asymptotically stable system, having distinct real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and singular values $\sigma_1, \sigma_2, \ldots, \sigma_n$ is searched, than it is necessary to solve a nonlinear algebraic equation system respect to the system residues. In this paper a numerical approach, founded on a nonlinear optimization algorithm, i. e. the simulated annealing algorithm, is proposed in order to find a suitable solution of the introduced nonlinear algebraic equation system and therefore to "solve" the problem of built-in systems with assigned invariants quantities. The goodness of the proposed approach is tested by considering some examples.

II - PRELIMINARIES

In this section some mathematical preliminaries are presented.

Theorem [3]:

Let us consider an asymptotically stable SISO system S ($A \in R^{nxn}$, B, $C^T \in R^{nx1}$) in minimal form, for such a system it is possible to introduce the Cross-Gramian equation as it follows:

$$W_{co} \cdot A + A \cdot W_{co} = -B \cdot C \qquad (1)$$

The eigenvalues μ_i (i=1,..,n) of the matrix W_{co} are real invariants of the system, whose modula are equal to the singular values σ_i (i=1,..,n). Moreover the sign of the W_{co} eigenvalues are other invariants that characterize the system signature matrix, described as it follows:

$$S = \mathrm{diag}\,(s_1, s_2, \ldots, s_i, \ldots, s_n), \qquad (2)$$

where $s_i = \pm 1$, since s_i is the generical sign of the W_{co} eigenvalue.

Proposition [1]:

Given an asymptotically stable system with distinct and real eigenvalues, the following relations hold:

$$\det(s_k \sigma_k \cdot I - R \cdot Q) = 0 \quad (k = 1, \ldots, n) \qquad (3)$$

where $R = \mathrm{diag}(r_1, r_2, \ldots, r_n)$ is a diagonal matrix containing the unknown system residues r_i, I is the identity matrix while each term of the Q matrix is given by the following expression:

$$q_{ij} = -(1/(\lambda_i + \lambda_j)) \qquad (4)$$

λ_i being the generic eigenvalue of the system.
The algebraic equation system (3) relates the system residues with its structural parameters; then, if at least a set values R_i can be found, the proposed problem can be solved and it is possible to assign both system structural parameters, e.g. eigenvalues and singular values.
An analogous formalization of the problem can be introduced in the case of placement of eigenvalues and characteristic values by using the Cross-Riccati equation studied in [4]. The equations (3) lead to a nonlinear algebraic equation system of the form:

$$F(\sigma_y) = \sum_{j=1}^{n} \binom{n}{n-j+1} \sum_{i=1}^{} F_{ij}(\sigma_y, Q) \cdot (R_{s_1} \cdot R_{s_2} \cdots R_{s_{n-j+1}}) = 0, \qquad (y1, y2, \ldots, n)$$

where $s_1, s_2, \ldots, s_{n-j+1}$ represent all the combinations without repetition of n-j+1 integers. The solution of this system depends on the matrix signature.

III - OPTIMIZATION STRATEGY

Referring to equation (3), our goal can be stated as it follows:

given s_i, σ_i, λ_i (i=1,..,n), let us determine the set of system residues $R_i = (r_1, r_2, \ldots, r_n)$,

belonging to an assigned domain, such that the R·Q matrix has eigenvalues θ_i as close as possible to the desired invariants $\mu_i = s_i \cdot \sigma_i$. Therefore the cost-function, that must be minimized, has been stated as it follows:

$$E(R_i) = \sum_{k=1}^{n} \left(\theta_k - s_k \cdot \sigma_k \right)^2 \qquad (5)$$

where:

θ_k is the k-th eigenvalue of the R·Q matrix corresponding to each considered $\tilde{R}_i = (r_1, r_2, .., r_n)$ set;

s_k and σ_k are respectively the k-th imposed signature and singular value.

The cost-function absolute minimum is known to be zero only in the case that the original problem admits a solution; when it does not exist we hope to find the system that guarantees the global minimum of the cost-function (5). To minimize $E(R_i)$ the Simulated Annealing optimization algorithm [5] has been adopted. The use of the "Annealing Algorithm" is due to avoid local minima that represent a heavy drawback in our search; in fact, differently from the classical optimization algorithms, the proposed one accepts not only those solutions that produce negative changes ΔE in the cost-function, but are taken into account also the ones giving positive changes that are accepted with a probability P given by:

$$P = \exp(-\Delta E/T),$$

where T is a control parameter called "effective temperature". From this latter consideration it follows the algorithm ability to go trough the local minima looking for a better result.
Let observe that the stated cost-function gives the square of the euclidean distance between the desired set of singular values and the found one. For each choice of the system eigenvalues, signature and singular value sets, no a priori condition ensures us about the solution existence, however the minimum of the cost-function found by the optimization algorithm gives us the system that best approaches the desired one.

IV - EXAMPLES

IV.1 - Example 1

Let us try to synthesize a system with the following poles, singular values and signatures:

λ= [1000 550 120 100 50 10]

σ=[308.5 69.9 8.6 0.2 0.17 0.02]

s=[-1 1 -1 1 -1 1]

By applying the simulated annealing in order to minimize the cost-function (5), the following residues R^* have been found:

$$R^* = [6501.6 \ -17875.4 \ -5228.7 \ 4611.6 \ -36652.3 \ 2624.2]$$

Moreover the equation (5) assumes the final value $E_{min}=10^{-4}$. Such a value implies that the stated problem allows a solution.

IV.2 - Example 2

In this second example we search for a system with the following eigenvalues, singular values and signatures:

λ = - [9 7 5 3 2 1]

σ = [90 80 70 60 50 40]

s = [-1 1 -1 1 -1 1].

Due to the fact that the cost-function value found in this case is $E_{min}=7059.9$, it is possible to assert that a solution, to the stated problem, does not exist. However the simulated annealing furnishes the system closest to the searched one.
The optimization algorithm gives the system characterized by the assigned eigenvalues and with the following residues:

$$R^* = [-166021.0 \ 299194.4 \ -178760.344 \ 50005.7 \ -11435.3 \ 426.9].$$

VI - CONCLUSIONS

In this paper it has been proposed a numerical solution of the problem of assigning system structural properties as eigenvalues and singular values. The employed optimization algorithm is based on the simulated annealing strategy.
The examples reported show the effectiveness of the proposed approach.
The procedure could be introduced for assigning other sets of invarants such as, for example, the eigenvalues and the characteristic values.

REFERENCES

[1] L.Fortuna, A.Gallo, G.Nunnari, "On Built-in Balanced Models With Guaranteed Structural Parameters", SIAM 1988 Annual Meeting, Minneapolis, Minnesota, July 1988.

[2] W.Gawronski, F.Y.hadaegh, "Balanced input-output assignment", Proc. of 28th Conf. on Decis. and Control, Tampa, Florida, December 1989.

[3] K.V.Fernando, H.Nicholson, "On the structure of balanced and other principal representations of SISO systems", IEEE Trans. on Autom. Control, AC-28 (1983), pp. 228-331.

[4] L. Fortuna, A. Gallo, G. Nunnari, "A new repersentation of SISO systems for studying approximated models", J. of Franklin Inst., 1988, 325, pp. 143-153.

[5] N.Metropolis, A.Rosenbluth, M.Rosenbluth, E.Teller, Journal Chem. Phys 21, 1087, 1953.

ESTIMATION OF PIECEWISE CONSTANT ROTATIONAL MOTION IN THE PLANE

by Coert Olmsted

Alaska SAR Facility, Geophysical Institute, University of Alaska Fairbanks

Abstract. Sea ice motion applications raise the problem of fitting a piecewise constant rotational model to a set of displacement vectors. The problem is twofold. Estimating a rotation from data and partitioning the data into rotationally constant classes.

To solve this image classification problem we use a reciprocal polar transformation and a linear clustering algorithm.

Introduction. Remote sensing of sea ice describes its motion by a two dimensional displacement field [Kwok et al. 1990]. Inspection (see Figure 1) supports the assumption of discrete elements, each of which rotates rigidly in the plane about a fixed center with a constant angular velocity [Olmsted 1991]. The model is rotationally piecewise constant with curvilinear discontinuities. The motion of each element can be specified by three parameters. the coordinates of the center and the angular speed.

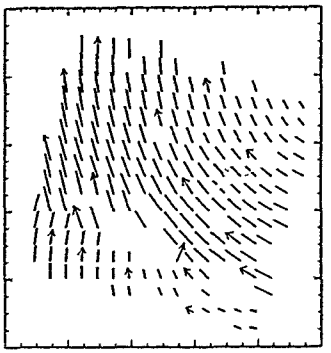


Figure 1. Displacement field for ice motion in the Beaufort Sea between SEASAT revolutions 1439 and 1482. Arrows on some vectors show the direction. The remaining displacements follow similar streamlines. The tails are on a regular 5 km grid.

The Problem. We need to estimate the parameters from a set of rotational displacement data with errors. This problem has been well studied [Watson 1988]. In our case it is complicated by the existence of several rotations the number of which is also an unknown. We must first partition the data into a number of elements, then estimate the constant rotation of each. The latter problem is relatively easy (See Appendix I). The former is a more difficult matter of image classification. A measure of the goodness of the classification is given by the total variation from constancy of all classes.

Brute force optimization of the total variation over all partitions is clearly too large a search. It is trivial that the optimal partition will be the discrete one in which each data point is a separate class. Then the total variation is zero but no information has been gained. Restricting the number m of classes still leaves non-contiguous elements. Thus only geometrical dissections are acceptable. This suggests using cluster analysis in some space including the plane of the motion.

Analysis. For instance in the case $m = 2$ we can consider partitions resulting from bisections of the plane by straight lines Parameterize a line with two real numbers (s,t), say slope and intercept. Then for each such pair the data will be divided into two sets. The sums of squares of deviations from estimated best constant rotation may be calculated for each set and then optimized with respect to s and t [Thorndike 1989]. This is a discrete objective function which jumps when the bisecting line crosses the location of a data point. Each additional bisecting line adds 2 dimensions to the search space, greatly increasing the complexity. This model is also restricted to straight line discontinuities.

Applying cluster analysis directly to the data space encounters the difficulty that a single displacement datum $(\Delta x, \Delta y)$ at (x,y) does not uniquely determine a rotation. Any pair of data determines a center of rotation at the intersection of the perpendicular bisectors to the displacement segments. However, the displacements may not be compatible with the center. To insure compatibility, cluster only intersections (x_0,y_0) of pairs of bisectors for data for which the displacements are roughly equal, i e., $\alpha_1 \approx \alpha_2$, where

$$\alpha_i := \frac{\sqrt{\Delta x_i^2 + \Delta y_i^2}}{\sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}}.$$

There are $n(n+1)/2$ pairs implying quadratic complexity. Then cluster in the velocity 3 space (x_0, y_0, α). To use contiguity information, however, we must include the coordinates of each datum, which raises the dimension of the search space to 5, an impractical level for the amount of data and compute power available.

Alternatively associate each perpendicular bisector with a point in such a way that images of lines which intersect at a common point have a distinguished configuration in the image plane. One such method is *reciprocal polars* [Bracewell 1990]. Parameterizing lines by slope and intercept, (s,t), note that if a family of lines are concurrent at (x_0,y_0), then for all s and t we have $y_0 = sx_0 + t$, i e., s and t are linearly related and the slope and intercept of that line in the s-t plane depend on (x_0,y_0) uniquely. This relationship is preserved by any affine transform of the s-t plane. Thus reciprocal polar points of perpendicular bisectors of data from distinct rotations will cluster along distinct lines in the s-t plane and this plane may be scaled and sheared so as to enhance the separation of the lines. Figure 2 shows the reciprocal polars of the data of Figure 1

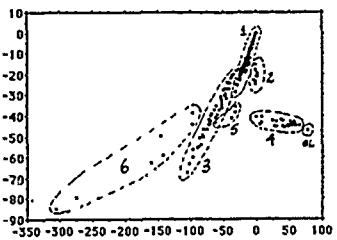


Figure 2. Scatter plot of the reciprocal polar points of all perpendicular bisectors of displacements from Figure 1 These fall into six natural groups as labeled 1 through 6 Where geometrical clustering is ambiguous, additional grouping information can be got by considering contiguity in the vector field, and the size of the displacements. Thus the outlier (marked OL) on the far right is compatible with the center of Group 4, but its magnitude is too large. Similarly the member of Group 5 whose reciprocal polar clusters with Group 6 is too small for that displacement. The units on the axes are pixel numbers scaled for clarity of the plot.

Given a cluster, linear regression will estimate the center coordinates. We must now, however, search for elongated clusters rather than round ones, a nonstandard image analysis problem. One might hope to find a transformation of the plane which would map distinct lines onto distinct points and thus simplify the clustering. It is elementary, however, to show that the only transformations of the plane which map all lines onto points are the constant ones. Thus there is no transform which injects lines onto points and we are stuck with elongated clusters. This problem can be attacked with specialized algorithms such as in Appendix II. The result is a determination of sets of displacement segments which have a common center as shown in Figure 3.

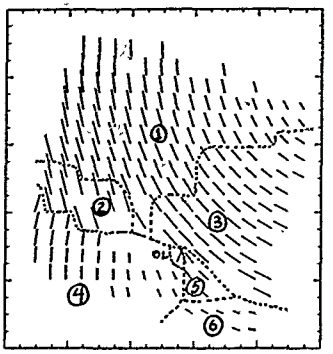


Figure 3. The vectors of Figure 1 grouped by dotted lines into the clusters of Figure 2 and labeled with corresponding numerals in circles.

APPENDIX I

For a set of displacements $\vec{x}_i \to \vec{y}_i$ in the plane, assume they are the result of a rotation $A(\theta)$ sampled with error and estimate the center $\vec{p} := (x_0, y_0)$ and angular displacement θ.

The first of two regression methods minimizes the total square error,

$$S(x_0, y_0, \theta) := \sum_i \| \vec{y}_i - \vec{p} - A(\vec{x}_i - \vec{p}) \|^2,$$

by setting the three partial derivatives of S to zero. This results in a 3×3 system of equations which is quadratic in $\cot \frac{\theta}{2}$ with explicit solution $\cot \frac{\theta}{2} = \beta \pm \sqrt{1+\beta^2}$ where

$$\beta := \frac{c_{x_1 x_2} + c_{y_1 y_2}}{c_{x_1 y_2} - c_{x_2 y_1}}.$$

The terms defining β are the covariances between the various coordinates of the displacement endpoints $\vec{x}_i =: (x_{1i}, y_{1i})$, $\vec{y}_i =: (x_{2i}, y_{2i})$. The center coordinates are then found in terms of the centroid of the displacement endpoints, the mean displacement lengths and $\cot \frac{\theta}{2}$.

The second method estimates the center as the point with least total square distance from the perpendicular bisectors of the displacement segments. If the perpendicular bisectors are parameterized as lines $c_i x + s_i y = p_i$, $c_i^2 + s_i^2 = 1$, then the total square distance is $S(x_0, y_0) = \sum_i (c_i x_0 + s_i y_0 - p_i)^2$. Setting derivatives to zero and solving for (x_0, y_0) gives

$$x_0 = (\textstyle\sum s_i^2 \sum p_i c_i - \sum s_i c_i \sum p_i s_i)/D$$
$$y_0 = (\textstyle\sum c_i^2 \sum p_i s_i - \sum s_i c_i \sum p_i c_i)/D$$

where $D = \sum s_i^2 \sum c_i^2 - (\sum s_i c_i)^2 = \sum_{i<j} \sin^2 \gamma_{ij}$ and γ_{ij} is the angle between lines i and j. The angular displacement θ can then be estimated as the mean of the angular displacements θ_i of each data displacement around the center (x_0, y_0).

Both of these methods appear to be numerically stable for small angular displacements even for the case of distant centers where the displacements are nearly parallel.

APPENDIX II

For a set of data points in the plane $S = \{P_i := (x_i, y_i)\}_{i=1}^n$ which are distributed with random error along a set of lines $\{y = a_i x + b_i\}_{i=1}^N$, estimate N and a_i, b_i.

Rotating Swath Algorithm

1. Translate the origin to $(x_m, y_m) \in S$, a point closest to the centroid of S
2. For a given half width w and a slope angle ϕ, let $F(w, \phi)$ be the number of data points falling within the swath of width $2w$ and slope $\tan\phi$ centered at (x_m, y_m). Compute this function as a two dimensional array for $\phi = 0, 175°, 5°$ and $w = \Delta w, W/4, \Delta w$ where $W := \text{diam} S$ and $\Delta w \approx W/50$.
3. Apply discrete optimization to this array of integers to find local maxima of F. One such should occur fairly near the low width side of the w, ϕ rectangle for each linear cluster of points.
4. If there is a distinct global maximum at, say (w_M, ϕ_M), let S_M be the set of points falling within this swath (of width $2w_M$ and slope $\tan\phi_M$ centered at (x_m, y_m)). Consider it as a seed population for the dominant linear feature of the data set.
5. Do linear regression on S_M to get a slope and intercept of the seed population cluster line. Add to this population all points of S for which the deviation from the line does not exceed the maximum deviation of the seed population. Recompute the linear regression on the expanded population and discard any outliers. Then iterate from the expansion step until no new points adhere to the cluster
6. Remove the cluster obtained in Step 5 from the general population. Then start over at Step 1. Iterate until all points are exhausted or no new clusters are produced.
7. To check, vary the initial point (x_m, y_m) and see if the same clusters appear.

REFERENCES

Bracewell, R. N., Reciprocal polars in the plane, *Personal communication*, 1990.

Kwok, R., J. C Curlander, R. McConnell and S S Pang, An ice-motion tracking system at the Alaska SAR Facility, *IEEE J. Ocean Eng.*, 15, 44-54, 1990.

Olmsted, C., Measuring sea ice deformation with imaging RADAR satellites, *Proceedings of the International Conference on the Role of Polar Regions in Global Change, June 1990*, Ed. G. Weller, University of Alaska Fairbanks, (To appear 1991).

Thorndike, A. S., *Personal communication on piecewise constant regression*, 1989.

Watson, G. S., Statistics of rotations, *Probability on Groups IX*, Ed. Heyer, Springer-Verlag, pp 398-413, 1988.

SIMULATION AND OPTIMIZATION OF AN AGGREGATES PRODUCTION INSTALLATION

JY. MONTEAU, M. LIU
SITIA, 39 avenue de l'Etrier, 44300 NANTES FRANCE

and

A. MALDONADO
LCPC, BP 19, Les Bauches du Désert, 44340 BOUGUENAIS, FRANCE

and

G. DELALANDE, M. CLEMENT
LRPC ANGERS, 23 avenue de l'Amiral Chauvin, 49130 LES PONTS DE CE, FRANCE

Abstract

This article describes the tranfer to an industrial site of quarry automation work done in the laboratory. It concerns the static simulation of a French quarry. The models of the devices are described first, together with the methods used to solve the problems inherent in the simulation of the processes (perturbations, search for equilibrium point). The simulation is described in detail, together with its use by an NLP program to find the optimum settings.

I - INTRODUCTION

Until the beginning of the 1980s, quarry automation work was limited to a combination of logical automation functions and local regulations. But in 1983 the LCPC [4] initiated a series of studies intended to optimize the operation of installations. Part of this work consisted of study of models of the equipments and their simulation [2], [3]. This article describes, on the occasion of the transfer of this work to an industrial site, the difficulties inherent in the simulation of this type of process and proposes solutions. The various points amplified are the description of the process, its modelling and simulation, and the search for the operating optimum by nonlinear programming (NLP), using the simulator.

II - THE PROCESS

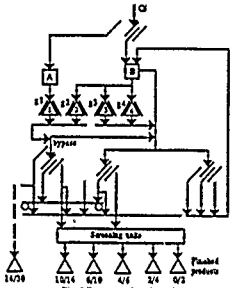


Fig. 1 Diagram of the installation

The installation consists of buffer silos, crushers, screens, and conveyors. By acting on the number of devices in operation and the routings, several possible operating diagrams can be obtained. One important characteristic of this system is the recycling to silo B of a part of the ouput of the three first screens : at the equilibrium regime, the levels of the silos are constant, and a variation of the setting of crushers 2, 3, and 4 can make the level of silo B unstable. The control conditions of this system are : the infeed flowrate Qi and the gaps of crushers g1, g2, g3 and g4. The output conditions are the finished-product flowrates $Q_{0/2}$, $Q_{2/4}$, $Q_{4/6}$, $Q_{6/10}$, $Q_{10/14}$. The perturbations are variations of grading and type of infeed material and of wear of the devices.

III - MODELLING AND IDENTIFICATION

The models of the devices are static models. Since the influence of the conveyors takes the form of simple pure delays, they have not been modelled.

3-1 Model of crusher

The perturbations are the origin not only of the instantaneous errors on the outputs predicted by the model, but also of the non-stationarities of the system. The former, not mesurable, are caused by the materials, and the latter by the metallurgical qualities of the wearing parts, revealed by a decline in the measured power over time. It is difficult to incorporate an estimate of wear, which varies condiderably from one set of parts to another, in the model. It affects the value of the operating point (W, Q). This is what led to a model based on small differences. The tests consisted of setting to an operating point, increasing and decreasing the clearances of the devices, and recording the outfeed curves. This gives linear functions.

The cruscher model, derived from the results of the campaign of tests, is a linear variation model.

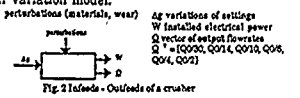


Fig. 2 Infeeds - Outfeeds of a crusher

with : $W = W_0 + \Delta W$

$Q = Q_0 + \Delta Q$

W_0 and Q_0 are the values of W and Q at the operating point, ΔW and ΔQ their variations such that

$$\Delta W = s_1 \Delta\varepsilon \qquad \Delta Q = s_2 \Delta\varepsilon$$

with $s^T_2 = (s_{0/2}, s_{2/4}, s_{4/6}, s_{6/10}, s_{10/14})$

To ensure the coherence of the models, the following are imposed :

$W \geq W^*$ $\qquad W = \sup(W^*, W_0 + \Delta W)$

$Q_{0/2} \geq 0$ $\qquad Q_{0/2} = \sup(0, Q_{0/2} + s_{0/2}\Delta\varepsilon)$

$Q_{0/4} \geq Q_{0/2}$ or $Q_{0/4} = \sup(Q_{0/2}, Q_{0/4} + s_{0/4}\Delta\varepsilon)$

$\vdots$ $\qquad \vdots$

$Q_{0/30} \geq Q_{0/14}$ $\qquad Q_{0/2} = \sup(0, Q_{0/14}, Q_{0/30} + s_{0/30}\Delta\varepsilon)$

with W^* the no-load power of the crusher.

The identification is easy : the outfeed values are recorded for two dif ferent infeed values.

$\varepsilon_1 \Rightarrow (W_1, Q_1)$ $\qquad \varepsilon_2 \Rightarrow (W_2, Q_2)$

From this are deduced : $s_1 = \dfrac{W_1 - W_2}{\varepsilon_1 - \varepsilon_2}$ $\qquad s_2 = \dfrac{1}{\varepsilon_1 - \varepsilon_2}(Q_1 - Q_2)$

3-2 Model of screen

Whithen's knowledge model (1972) [1] was used . the probability of a particle of size x not passing through a screen of mesh width h and wire diameter d is considered :

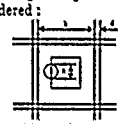


Fig. 3 Diagram of a screen mesh

The parameter m represents the effectiveness of screening , it de pends on the infeed flowrate. Function C (x) is called the classication function. Let $S^T = [s_1, s_2, \ldots s_n]$ be the vector of the particle sizes. Here $S^T = [30\ 14\ 10\ 6\ 4\ 2]$. The mean probability for a particle to be oversize is calculated by :

$$Ci = \frac{1}{S_i - S_{i+1}} \int_{S_{i+1}}^{S_i} C(x)\,dx \qquad \text{for i from 1 to n-1}$$

$$Cn = \frac{1}{S_n} \int_{0}^{S_n} C(x)\,dx$$

This defines a diagonal matrix $C = \begin{bmatrix} C1 & & \\ & C2 & \\ & & CN \end{bmatrix}$

This matrix may be used to calculate the throughtputs of materials in each grading class for oversize and undersize.

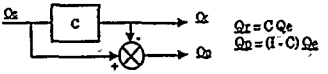


Fig. 4 Schematic diagram of screen model

Where I is the identity matrix and Qe, Qr, Qn the infeed, oversize and undersize throughput vectors. In the absence of precise measurements, the parameter m has been taken equal to 20, a realistic value

IV - THE SIMULATION

The system consists of a set of machines interconnected by conveyors. The output variables of one constitute the inputs of another. They are calculated using models of the devices in the order upstream-downstream. However, because of the recycling to silo B (see Fig. 1), the output of crushers 2, 3, and 4 acts on their own infeed. This is why the calculation requires several iterations until the power, flowrate, and grading variables cease to change as judged by some criterion. There is therefore a fundamental problem here in the simulation of these processes, the problem of convergence, a problem that is still unsolved. The steady state has been simulated as shown by the following flow sheet (Fig. 5), usable in the general case (the devices are numbered in upstream-downstream order).

For a particular set of control values (g1, g2, g3, g4, Q1), in the general case, the infeed flowrates of the silos are not equal to their output flowrates, and the levels, are unstable. To locate an equilibrium point (level of silos constant), the settings of crusher 1 and the infeed flowrate of the installation are calculated as a function of the settings of crushers 2, 3, and 4 by an iterative method. This problem of equilibrium in a process involving flows of materials is common. The method of attaining equilibrium used here is shown in figure 6.

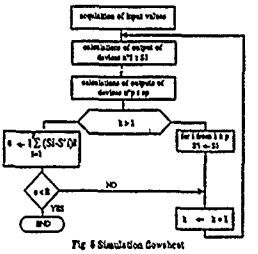


Fig. 5 Simulation flowsheet

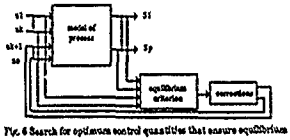


Fig. 6 Search for optimum control quantities that ensure equilibrium

V - OPTIMIZATION

Software for optimization of the installation using the simulator has been developed to determine the optimum settings to be used to maximize or minimize the production of one class of aggregates or another.

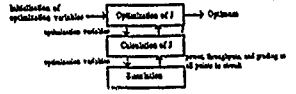


Fig. 7 Search for optimal operating point

The optimization, by nonlinear programming, uses an augmented Lagrangian method. The Lagrangian is minimized without constraint by a method of Newton The criterion maximized is the sum of the throughputs of the finished products. Contraints are used to favour one type of output or another : lower and upper limits are imposed on the throughputs of each finished product.

$$\text{e.g.: } J = \frac{1}{Q10/14 + \ldots + Q02} + \frac{1}{Q} \quad \alpha \begin{cases} 30\% \le 10/14 \le 40\% \\ \vdots \\ 0\% \le 0/2 \le 5\% \end{cases}$$

The optimization variables are the settings of crushers 2, 3 , and 4

VI - EXAMPLE

The user has here sought to obtain mainly 10/14 and 6/10 while minimizing the quantities of 2/4 and 0/2 produced (crushers 525, 535, 625, 645 = crusher 3, 4, 2, 1 - D = 30 mm).

Iteration number : 224.
Clearances (discrete values) 525, 535, 625, 645 = 5 -2 -4 -5.
Infeed flowrate : 743 Tph, required flowrate : 1 to 1000 Tph.
Throughputs of crushers 525, 535, 625 645 (Tph) = 241 273 282 584
Powers of crushers 525, 535, 625 645, (KW) = 280 166 89 133

Crushers gradings (%)

	0/D	0/14	0/10	0/6	0/4	0/2
G625	100.0	76 2	56.0	43.2	39 5	32.8
G635	100.0	71.2	52.1	40.8	37.5	29.0
G625	100.0	98 5	51.6	21.2	13 0	5.1
G645	100 0	53 8	38 0	22.3	14.2	6 5

Finished products

Class	Throughput (Tph)	Percentage (%)	Percentage desired (%)
10/14	361	49	30 to 50
6/10	171	23	30 to 50
4/6	78	10	10 to 20
2/4	74	10	5 to 10
0/2	60	8	0 to 5

VII - CONCLUSION

In this article, on the occasion of the work done on an industrial case, we have attempted to show the difficulties inherent in the simulation of quarry installations and to propose some solutions.

The main difficulties encountered are in taking into account perturbations in the materials, the modelling, and, in the case of a static simulation, the search for settings of the devices such that the flows of materials are in equilibrium, i.e. the levels of buffer stocks are constant.

The first difficulty was resolved by using models of a variation about an operating point and the second by the use of a corrector calculating the settings on the basis of a criterion of equilibrium.

This simulation work is only a first step towards the automation of this type of process.

REFERENCES

[1] W.J. WHITTEN, "The simulation of crushing plants with models developped using multiple spline regression", Journal of the South African Institute of Mining and Metallurgy, pp 22-263, May 1972.

[2] J.Y MONTEAU, "Contribution to the automation of crushing-screening installations. Modelling of the devices. Static simulation", Doctoral thesis in Engineering, University of Nantes, 1987 (in French).

[3] M. LIU, "Decoupling of systems with delay , Simulation and optimal control of an aggregates production installation", Doctoral thesis, Univertity of Nantes, 1988 (in French).

[4] S. LEBAS, A. MALDONADO, "A short testing centre serving research and its promotion in the quarrying industry", Bulletin de Liaison, Laboratoire Central des Ponts et Chaussées 160, February-March 1989, ref. 3361 (in French).

RESONANT MODES COMPUTATIONS IN DIELECTRIC-LOADED WAVEGUIDES WITH EGDE ELEMENTS

A. BOURHATTAS - L. PICHON
Laboratoire de Génie Electrique de Paris
U.R.A. D0127 CNRS
Ecole Supérieure d'Electricité
Universités Paris 6 et Paris 11
Plateau du Moulon
91192 Gif sur Yvette Cédex France

Abstract In this paper, resonant modes of different waveguides are computed with "edge elements" . These finite elements avoid all the "spurious modes", the non-physical numerical fields obtained from the solution of eigenvalue problems with classical finite element formulations. Moreover, comparisons with analytical results and previously published ones show the great accuracy of the numerical technique. Both empty and dielectric-loaded waveguides are analyzed.

I. INTRODUCTION

The finite element method seems to be a powerful tool for the analysis of arbitrary shaped waveguides : cut-off frequencies can be obtained as solutions from an eigenvalue problem. The most serious difficulty in these studies is that the computed solutions are plagued by non-physical (or "spurious") solutions : solutions which do not automatically satisfy the divergence free condition implied by the Maxwell's equations. Many attempts are performed to circumvent these unwanted fields [1]-[3]. It has been observed that discretized fields with continuous tangential components suppress the "spurious modes" [1] but no precise argument was put forward to explain the importance of this approximation. Recently, Bossavit showed the reason for which "edge elements" (a class of mixed finite elements [4] [5]) will not generate "spurious modes" [5]. Here, we have developed such a numerical approach [6] and applied the technique for the analysis of empty and dielectric-loaded waveguides.

II. VARIATIONAL FORMULATION

We deal with the Maxwell's time harmonic equations inside a 2-D bounded region Ω. This region contains lossless materials and is surrounded by an ideal conductor ($e \wedge n = 0$).

A weak formulation of the wave equation for the transverse electric field $e = (e_x, e_y)$ in Ω can easily be derived :

$$\int_\Omega \text{rote.rote}' \, d\Omega - k_c^2 \int_\Omega \epsilon_r \, e.\epsilon' \, d\Omega = 0 \quad (1)$$

where the values of k_c are the cut-off wavenumbers.

III. FINITE ELEMENT FORMULATION

The region Ω is discretized with a classical triangular mesh.

The edge elements have the following properties : the degrees of freedom e_a and the trial functions w_a are associated to the mesh edges ; e_a is the circulation of "e" along the edge "a" and w_a can be expressed as:

$$w_a = \lambda_i \, \text{grad}\lambda_j - \lambda_j \, \text{grad}\lambda_i \quad (2)$$

where λ_i and λ_j are the barycentric coordinates.

In each triangle, $e = (e_x, e_y)$ is :

$$e_x = \alpha_1 - \alpha_2 y \quad ; \quad e_y = \alpha_3 + \alpha_2 x \quad (3)$$

where $\alpha_1 \in R$; $\alpha_2 \in R$; $\alpha_3 \in R$

Substituing in (1) the discretized fields obtained with (3) for "e" and "e'" , we deduce a generalized algebraic eigenvalue problem of the form :

$$A \, u = k_c^2 \, B \, u \quad (4)$$

A (the "stiffness" matrix) and B (the "mass" matrix) have dimensions $n_a \times n_a$ where n_a is tne number of edges in the finite element mesh.

IV. NUMERICAL RESULTS

1. Empty waveguides

A hollow rectangular waveguide (1 x 2) was analyzed. The number of unknown edges is 126. This simple case gives spurious modes when solved with classical finite elements. [3]. Here, no spurious mode is observed. The first lowest modes were computed (table 1) and compared with analytical solutions. The relative error never exceeds 0.6%.

Modes	k_c	error(%)
TE_{10}	1.569	0.05
TE_{01}	3.140	0.03
TE_{20}	3.136	0.17
TE_{11}	3.506	0.18

Table 1 Computed k_c for a rectangular waveguide

A similar calculation was performed for a hollow circular waveguide (Radius :1) with 130 unknown edges. No spurious mode was observed. The results are shown on table 2.

Modes	k_c	error (%)
TE_{11}	1.860	1.
TE_{21}	3.091	1.2
TE_{02}	3.852	1.2
TE_{31}	4.273	1.7

Table 2 Computed k_c for a hollow waveguide

Figure 1 shows the distribution of the electric field for the TE_{11} mode.

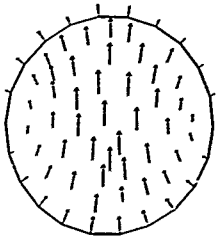


Figure 1 Electric field for the TE_{11} mode

2. Dielectric-loaded waveguide

A rectangular waveguide half-filled with dielectric material (figure 2) was studied.

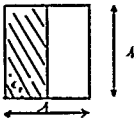


Figure 2 Half filled dielectric waveguide

This example is used in the literature as a test case for calculation methods. The number of unknown edges is 210. Spurious modes often arise in this problem [2]. Here, no one occured : all the computed solutions correspond to physical cut-off values. Table 3 exhibits the lowest computed values of k_c for different permittivities ϵ_r. A comparison of our results with those of a previous work [7] (obtained by means of truncated series) shows a good agreement.

ϵ_r	Our Method	Refer. [7]
2.	2.516	2.531
3.	2.240	2.161
4.	1.881	1.909
5.	1.698	1.729

Table 3 Lowest values of k_c for half-filled dielectric waveguide

V. CONCLUSION

"Edge elements" have been used to model empty and dielectric-loaded waveguides. The first cut-off frequencies were computed ; comparisons with analytical results values and with previously published ones show the efficiency of the method.

These elements avoid all the well known "spurious modes" and seem very promising for the study of more complicated problems in high frequency applications.

V. REFERENCES

[1] KIKUCHI F. "Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism" , Comp. Meth. Appl. Engng. , Vol 64 , pp 509-521 , 1987.

[2] HAYATA K., KOSHIBA M. , EGUCHI M., SUZUKI H. , "Vectorial finite-element method without any spurious solutions for dielectric waveguiding problems using transverse magnetic-field component", IEEE Transactions on Magnetics , Vol MTT-34 , n°11 , pp 1120-1124 , 1986.

[3] RAHMAN B.M.A., DAVIES J.B., "Penalty function improvement of waveguide solution by finite element", IEEE Transactions on Microwave Theory and techniques, Vol MTT-32, n°8, pp 922-928, 1984.

[4] NEDELEC J.C. , "Mixed finite elements in R^3 " , Numer. Math. , Vol 35 , pp 315-341 , 1980.

[5] BOSSAVIT A. , "Simplicial finite elements for scaterring problems in electromagnetism", Comp. Meth. Appl. Engng. , Vol 76 , pp 299-316 , 1989 .

[6] PICHON L.,RAZEK A., "3-D cavity resonances computations without spurious modes", 4th International IGTE Symposium, Gratz, Austria , 10-12 October 1990.

[7] MA J. G., "Numerical analysis of the characteristics of TE-modes of waveguides loaded with inhomogeneous dielectrics", Proc IEE Pt H , Vol 138 , n°1, pp 109-112 , 1991.

AN APPLICATION OF THE EULERIAN-EULERIAN TECHNIQUE TO A TRANSIENT TWO-PHASE FIRE-SPRINKLER SIMULATION

N.A. BOTHAKIS AND E.R. GALEA

Centre of Numerical Modelling and Process Analysis
Thames Polytechnic
Wellington Street
LONDON SE18 6PF, UK

Abstract

This paper extends the field modelling approach to the simulation of enclosure fires to include fire-sprinkler interaction. The volume fraction or Eulerian-Eulerian approach is used to simulate the two phases. The hot fire gases represent one phase while the liquid water particles the second. The mathematical model presented takes into account the three modes of interaction between the phases: drag, heat- and mass-transfer. The resulting finite-difference equations are solved for using the commercially available computer package PHOENICS, which employs the two-phase algorithm IPSA (Inter-Phase Slip Algorithm). Comparisons between experimental and predicted results are presented. Current research is directed towards improving the efficiency of both the algorithms and their implementation.

Introduction

During 1989, fire claims cost the Association of British Insurers in excess of £600 million and the lives of about 1000 people. The installation of sprinkler systems within occupied enclosures has always been seen as an effective means of reducing fire losses. Fires in structures which are protected by sprinkler systems are often controlled and extinguished before the arrival of the local fire brigade resulting in minimal property damage. The loss of life is also greatly reduced [1,2].

In order to efficiently combat fire it is necessary to understand the nature of the interaction between the hot combustion products and the liquid water. This enables fire engineers to optimise the design and location of sprinkler devices.

For many years physical experimentation into these phenomena has been the main means of investigation. However, the amount of human and financial resources required to carry out a full-scale fire test with completely fitted enclosures, such as a room or aircraft fuselage, can be extremely expensive or even impossible. Additionally, it is not always possible to conduct enough fire tests to adequately deal with all alternatives, such as the nature and position of fire sources, ventilation alternatives and sprinkler options. Mathematical modelling provides a means to overcome these difficulties.

Mathematical Modelling

Mathematical modelling of fires and smoke spread in enclosures has received a considerable amount of attention. Field modelling represents the most sophisticated modelling strategy available for the simulation of enclosure fires. This deterministic approach [3,4] provides us with a clearer understanding of the complex processes occurring within a fire enclosure. Studies of different fires within a specified compartment geometry can be carried out with ease and with less expense than full-scale experiments. Field models involve the numerical solution of recirculating three-dimensional turbulent buoyant fluid flow with heat and mass transfer. As a result they consume a considerable amount of computer power.

Fire-sprinkler interaction also lends itself to this form of analysis. This problem is considerably more complicated than the straightforward fire in enclosure situation. There are now two physical phases which must be incorporated into the overall mathematical description. These are the gas phase involving the general fluid circulation of the hot combustion products and the liquid phase, representing the water droplets which have been injected into the fire compartment and evaporate. Previous studies by the authors were concerned with steady-state simulations [5,6]. Here, the analysis is extended to deal with the transient nature of the phenomena [6].

The Mathematical Description of the Physical Problem

Fire creates a strong buoyancy driven flow which gives rise to large-scale turbulent motion controlling the diffusion of mass and momentum. Water droplets injected into this hostile environment interact with the hot gases by being entrained into the thermal plume or by re-directing the flow of the hot gases.

The independent variables used to model this fire sprinkler interaction phenomena are the width, height and length: x, y and z of a cartesian co-ordinate system, as well as time t.

The 15 dependent variables requiring solution are the six velocity components (u_g, v_g, w_g) for the gas phase and (u_l, v_l, w_l) for the particulate phase in their respective cartesian direction (x,y,z), along with the pressure p, which is assumed to be the same for both phases. The enthalpies for the gas and water phases, h_g and h_l respectively, along with the concentration of water vapour within the gas phase, c. The gas and liquid volume fractions r_g and r_l are solved for including the effect of evaporation. The 'shadow' volume fraction, r_s, is the volume fraction in the absence of evaporation. Finally the turbulence kinetic energy and dissipation rate of the gaseous phase (k,ε). Turbulence in the liquid phase is neglected. The shadow volume fraction technique allows us to evaluate the diminishing droplet size during evaporation [7].

Governing Differential Equations

All these dependent variables, with the exception of pressure, appear as the subject of the differential equations of the form:-

$$\frac{\partial}{\partial t}(r_l\rho_l\phi_l) + \mathrm{div}\,(r_l\rho_l V_l \phi_l - r_l \Gamma_\phi \,\mathrm{grad}\,\phi) = r_l S_l \quad (1)$$

where ϕ stands for a general fluid-property such as velocity or enthalpy, Γ and S are diffusion coefficient and source terms respectively, and subscript i refers to the phase in question (gaseous or particulate).

The Mass-Conservation Equation

The pressure variable is associated with the continuity equation:-

$$\frac{\partial}{\partial t}(r_l\rho_l) + \mathrm{div}\,(r_l\rho_l V_l) = S_l \quad (2)$$

where ρ_i is the density of the phase.

The volume fractions are related to each other by the 'space' sharing condition:-

$$r_g + r_s = 1.0 \qquad (3)$$

The source/sink term S_i is the rate of evaporation (source for the gas, sink for the liquid).

Auxiliary Equations

Due to the very nature of the problem, i.e. the interaction between two phases, certain correlations need to be included in the model to close the problem. These time-dependent relations deal with the interphase heat and mass transfer and the friction between the gas and the liquid phases. The assumptions made are that the gas and droplets are dispersed within the control-volume and that the droplets are spherical. The last assumption is not essential but simplifies the nature of the empirical input [6].

The Solution Procedure

The above equations are solved for using the computer program PHOENICS [8] with the built-in equations solvers SIMPLEST and IPSA (Inter-Phase Slip Algorithm). The latter, a more elaborate solution procedure which is used for multi-phase modelling, is able to handle the presence of two simultaneously present phases sharing a common pressure. It evaluates the increased number of governing equations of the flow and the strong interaction between them, such as interphase friction and mass, as well as the space sharing of the volume fractions condition. For these calculations PHOENICS uses the conventional staggered grid approach for solving finite-volume equations [8].

Fire-Sprinkler Simulation

In order to evaluate the validity of the model outlined above, it is necessary to compare experimental fire-sprinkler data with model results. Two fire scenarios have been investigated. The first concerned a corner fire located in an OPEN office-sized compartment [6,9]. The simulation results agreed reasonably well with the small amount of experimental data available for comparison purposes. The results presented here concern a bed fire within a large CLOSED hospital ward. The experimental results were produced by the U.K. Fire Research Station [10].

The fire compartment depicted in Figure 1 was a six bed hospital room of dimensions 7.325m by 7.85m by 2.7m. The flow domain was fitted with a 14x15x11 cartesian grid comprising of 2310 control volumes. The thermal conductivity and thickness of the walls, floor and ceiling were supplied and included in the calculations. In addition, six 1kW oil filled radiators situated on a wall were modelled using a single 6kW heat source 7.85m long, 0.3 in width and 0.5m high.

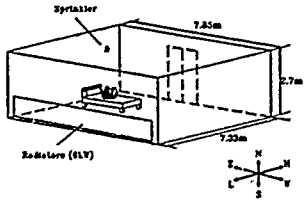


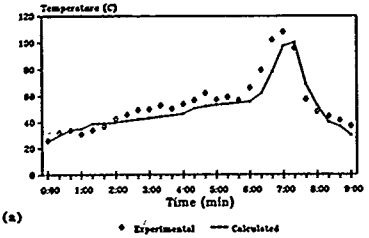
Figure 1 Hospital Room Schematic

In the experiment the fire was situated on a bed which was 1.35m wide, 1.75m in length and stood 0.5m above the floor. The bed-head was positioned at the centre of the east wall. The sprinkler was located 0.35m below the ceiling and 0.895m from the east wall along the centre line of the room. The water was released at a rate of $0.566 \times 10^{-3} m^3/s$ at an angle of 70° from the sprinkler heads axis of symmetry. This dispersion pattern took into account the sprinkler's deflector plate. It was assumed that droplets with a uniform average diameter of 1mm were released with an initial temperature of 10°C.

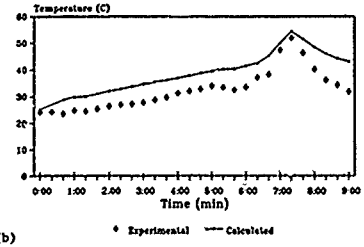
The room conditions prior to the sprinkler activation were obtained by modelling only the fire. The results from the single-phase simulation, which lasted for 420 seconds, were used as initial values for the two-phase fire-sprinkler study. During the single-phase simulation the fire was ramped from 0kW to 40kW according to experimental data. The fire was modelled as a heat source occupying a fixed area of 0.45mx0.5m, positioned 0.25m from the east wall. After 420 seconds the sprinkler was activated. During this stage, which lasted for 120 seconds, the fire was maintained at a constant 40kW. The simulations were performed with a 1 second time step.

Results

During the experiment gas temperatures were monitored at seven main locations. The thermocouples were positioned 75mm below the ceiling. Comparisons between experimental and numerical gas temperatures for two sites are presented in figure 2. The thermocouples were located between the heat source and the east wall (figure 2-a) and in the vicinity of the room's west end high corner (figure 2-b). These clearly indicate that the correct trends in temperature variation have been captured. Experimental and numerical uncertainties concerned with the fire load and the coarseness of the numerical grid contribute towards the observed discrepancies.



(a)



(b)

Figure 2 Predicted and measured gas temps at two thermocouple locations.
a)located between heat source and east wall
b)located in the vicinity of west end high corner

The following three diagrams are taken through the centre of the room and through the fire and sprinkler sources. Figures 3 and 4 show gas temperatures and velocity vectors before and 120 seconds after sprinkler activation. These clearly show the manner in which the sprinkler creates a water curtain confining the hot gases to a small area of the room. Eventually the gases remote from the fire and sprinkler are cooled down. As can be seen in figure 4-a, prior to sprinkler activation there exists one large re-circulation cell driven by the heat source. Two minutes after sprinkler activation (figure 4-b) two major flows are apparent. The first, generated by the sprinkler is downwards, deflecting off the bed and along the floor. The second, generated by the fire, is along the ceiling. These two current meet towards the centre of the room, aiding the mixing and cooling process.

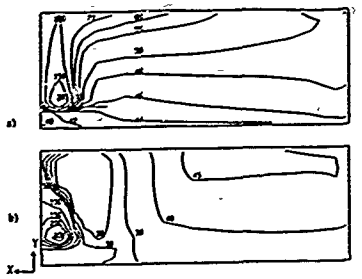


Figure 3: Predicted Gas Temp.Contours
a)fire only (420 secs) b)fire and sprinkler (540 secs)

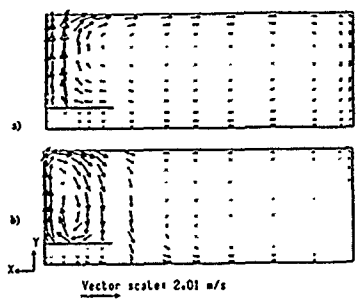


Figure 4: Predicted gas velocity vectors
a)fire only (420 secs) b)fire and sprinkler (540 secs)

Finally the spread of the water droplets in terms of their volume fractions are shown in figure 5 after 120 seconds.

A further observation to emerge from these simulations is that this approach is very expensive in terms of computing time. The simulations were performed on a Norsk Data ND-5900 machine, roughly the equivalent to a VAX 11/780. The calculations required 75 hours of CPU time to simulate the 420 seconds prior to sprinkler activation. The next 120 seconds consumed 216 hours. These calculations were performed on a relatively coarse grid. Clearly if this technique is to be adopted as an engineering design tool a means must be found of reducing this enormous computing effort. Parallel computing techniques offer a way of achieving this. At

Figure 5. Predicted volume fractions of water phase

Thames Polytechnic we have modified the fluid flow package HARWELL-FLOW3D [11] to make efficient use of the multi-processor architecture offered by the INMOS Transputer. Early work reveals that on test problems involving up to 17,640 computational cells and utilising 15 transputers, efficiency of over 86% can be achieved. On a 15 processor transputer system this results in a 13 fold speed-up [12]. Translating this performance to the above fire-sprinkler simulation and using 15 transputers it is expected that the 291 hour simulation could be performed in about 22 hours.

Conclusions

The results presented above indicate that time-dependent fire-sprinkler scenarios can be simulated using the outlined volume fraction approach. Further studies are currently in progress to investigate the effect of flow rate and droplet size. Furthermore, detailed grid-refinement studies need to be undertaken. It is envisaged that these and other computationally intensive fire field modelling calculations will be pursued on parallel architecture machines.

Acknowledgement

The authors would like to thank Professor N.C. Markatos for many useful discussions in the early stages of this project, CHAM for allowing the use of PHOENICS, as well as the SERC and Ove Arup and Partners for funding this project.

References

1. T.Z. Harmathy,"On the Economics of Mandatory Sprinklering of Dwellings", Fire Tech.,1988, pp.245-261
2. Anon, "Fires controlled by Sprinklers", Fire Prevention, Vol.226, 1990, pp.48-49
3. E.R. Galea, "On the Field Modelling Approach to the Simulation of Enclosure Fires", J. of Fire Prot. Eng., Vol.1(1), 1989, pp.11-22
4. E.R. Galea and N.C. Markatos, "The Mathematical Modelling and Computer Simulations of Fire Development in Aircraft", Int. J. Heat Mass Transfer, In press
5. N.A. Hoffmann, E.R. Galea and N.C. Markatos, "Mathematical Modelling of Fire Sprinkler Systems", Appl. Math. Modelling, Vol.13, 1989, pp.298-306
6. N.A. Hoffmann, Computer Simulation of Fire-Sprinkler Interaction, PhD Thesis, Thames Polytechnic, London, 1990
7. D.B. Spalding, "The shadow method of particle-size calculation in Two-Phase Combustion", 19th Symp. on Combustion, The Combustion Inst., 1982, pp.941-951
8. D.B. Spalding, "A General-Purpose Computer Program for Multi-Dimensional One- Two-Phase Flow", Prepr. 81-6,IMACS, Vol. 23, 1981, pp.267-276
9. N.A Hoffmann and E.R. Galea, "A Transient Two-Phase Fire-Sprinkler Simulation", presented at the 5th Conf. ECMI, Lahti, Finland, June 1990
10. P.G. Smith, Private Communication, Fire Research Station, U.K.; 1987
11. A.D. Burns, I.P. Jones, J.R. Kightley and N.S Wilkes, "HARWELL-FLOW3D", User Manual, 1989
12. M. Cross and S. Johnson, "Mapping CFD algorithm onto fine grained parallel architecture", in prep.

NUMERICAL SIMULATION OF ENHANCED OIL RECOVERY

ROBERT H.J. GMELIG MEYLING

Royal Dutch/Shell

Exploration and Production Laboratory

P.O. Box 60 2280 AB Rijswijk The Netherlands

Abstract

Polymer flooding is an oil recovery technique, where polymer is added to water which is injected into an oil reservoir to increase hydrocarbon recovery. Polymer flooding is a complex physical process, which can be described by a system of nonlinear partial differential equations. A numerical method for simulating polymer floods is presented. The method is based on special numerical techniques, which introduce almost no numerical diffusion and which permit relatively large time steps. Locally refined and dynamically adapted grids are used to accurately represent sharp, moving fluid interfaces and to reduce computational costs.

Introduction

Enhanced Oil Recovery (EOR) deals with cost-effective techniques to increase oil production from existing fields. An example of an EOR- technique is polymer flooding of fields containing heavy oil. The efficiency of the recovery process can greatly be improved by adding highly viscous polymer to the injected water. The polymer is costly and is typically injected in slugs.

Polymer flooding is a complicated process involving a fluid composed of oil, water and polymer. Water and oil are assumed to be immiscible, while polymer mixes only with water. The flow of fluid through a porous, heterogeneous reservoir is induced by the pressure gradient between injection and production wells and is affected by gravitational forces. The distribution of components (oil, water, polymer) over the reservoir is governed by nonlinear convection (shock fronts), physical diffusion (capillary pressure), and adsorption of polymer by the rock. Fluid interfaces may become unstable due to large differences in viscosity of the different fluids. A fluid may form so-called *viscous fingers*, which channel through another fluid towards the production wells. This will clearly reduce the efficiency of the oil recovery process.

Accurate simulation of polymer floods is quite difficult but of vital importance to an oil company wishing to optimize the polymer injection strategy. Conventional (finite difference) methods often introduce excessive numerical diffusion, which badly smeares sharp fluid fronts and which destroys most of the dynamical behaviour of the simulated slug flow. Our solution method is based on decoupling the equations during timestepping and treating each equation by the most accurate and efficient numerical technique. All calculations are performed on locally refined grids in up to three space dimensions. The grids are dynamically adapted during simulation so that solution accuracy is enhanced and computational costs are reduced.

Mathematical equations of polymer flooding

A mathematical model describing polymer flooding of a porous medium consists of a coupled system of nonlinear partial differential equations. Let $\mathbf{x}$ denote the position in the reservoir and let t be time. Pressure $p(\mathbf{x},t)$ and velocity $\mathbf{u}(\mathbf{x},t)$ of the total (incompressible) fluid are governed by *Darcy's law*

$$\nabla.\mathbf{u} = q(\mathbf{x},t) \quad , \quad \mathbf{u} = -\mathcal{K}(\mathbf{x})\lambda(s,c)(\nabla p - \rho(s,c)\nabla d(\mathbf{x})) \quad (1)$$

with q the well flow rate; $\mathcal{K}$ the absolute rock permeability; λ the total mobility of the fluid; ρ a gravity term; and d the reservoir depth. The fluid consists of two immiscible phases: an *aqueous* phase containing water and polymer and an *oleic* phase containing only oil. Saturation $s(\mathbf{x},t)$ of the aqueous phase and concentration $c(\mathbf{x},t)$ of polymer in water satisfy the mass conservation equations

$$\phi(\mathbf{x})\frac{\partial s}{\partial t} + \nabla.(f(s,c)\mathbf{v}(\mathbf{x},t)) - \nabla.(\mathcal{D}^s(s,c)\nabla s) = q_s(\mathbf{x},t) \quad (2)$$

$$\phi(\mathbf{x})\frac{\partial (sc + a(c))}{\partial t} + \nabla.(cf(s,c)\mathbf{v}(\mathbf{x},t)) - \nabla.(\mathcal{D}^c(s,c)\nabla c) = cq_s(\mathbf{x},t) \quad (3)$$

with ϕ the rock porosity, $f(s,c)\mathbf{v}$ the aqueous phase velocity, q_s the water source term, and $a(c)$ the fraction of polymer adsorbed by the porous rock. The tensor $\mathcal{D}^s$ reflects diffusion at the oil/water interface due to capillary pressure effects, while $\mathcal{D}^c$ models (molecular) diffusion of polymer within water

Numerical simulation

The total fluid velocity $\mathbf{u}$ and pressure p vary slowly with time and are weakly coupled to saturation s and concentration c. Hence, a sequential approach is used to solve the equations Each timestep, the equations are separated into an *elliptic* equation (1) for velocity and pressure, and a system of convection-dominated diffusion equations (2,3) for s and c. Assuming s and c are known functions of $\mathbf{x}$, equation (1) is discretized using *mixed finite elements* and solved by an efficient *multigrid* technique (Schmidt and Jacobs, 1988) Operator-splitting is then applied to the system (2,3) to deal separately with the effects of *convection / adsorption*

$$\phi\frac{\partial s}{\partial t} + \nabla.(f(s,c)\mathbf{v}) = q_s \; , \; \phi\frac{\partial (sc + a(c))}{\partial t} + \nabla.(cf(s,c)\mathbf{v}) = cq_s \quad (4)$$

and *diffusion*

$$\phi\frac{\partial s}{\partial t} - \nabla.(\mathcal{D}^s(s,c)\nabla s) = 0 \; , \; \phi\frac{\partial (sc)}{\partial t} - \nabla.(\mathcal{D}^c(s,c)\nabla c) = 0 \quad (5)$$

The nonlinear system of *hyperbolic* equations (4) is solved by the *method of characteristics* combined with *Riemann solvers*. Nonlinearities in the system are governed by the *fractional flow* function $f(s,c) = \lambda_w/(\lambda_w + \lambda_o)$, where $\lambda_w(s,c) = s^2/\mu_w(c)$, $\lambda_o(s) = (1-s)^2/\mu_o$ are the aqueous and oleic phase mobilities and μ_w, μ_o are the viscosities of water (polymer-dependent) and oil. The function $f(s,c)$ is S-shaped (with a single inflection point) and has properties: $f(0,c) = 0$, $f(1,c) = 1$, $f_s(s,c) > 0$, $f_c(s,c) < 0$, for $0 < s < 1$, $0 \le c \le 1$. Adsorption of polymer by the rock satisfies the conditions: $a(0) = 0$, $a_c(c) > 0$, $a_{cc}(c) < 0$, for $0 < c < 1$. The velocity v can be obtained directly from the total fluid velocity u. The multi-dimensional problem (4) can thus be reduced to *one-dimensional* conservation laws

$$s_t + f(s,c)_\xi = q_w/\phi \quad , \quad (sc + a(c))_t + (cf(s,c))_\xi = cq_w/\phi \qquad (6)$$

along *streamlines*. These streamlines are defined as the integral curves $x(\xi)$ of the system of ordinary differential equations $dx/d\xi = v(x)/\phi(x)$. This system can be solved very efficiently by exploiting the mixed finite element representation of v.

If s and c are represented by constants on each grid block, the essential task is now to solve a sequence of (1-D) *Riemann problems* along streamlines. Writing (6) in quasi-linear form, each Riemann problem consists of the hyperbolic system

$$\begin{pmatrix} s \\ c \end{pmatrix}_t + \begin{pmatrix} f_s & f_c \\ 0 & f/(s+a_c) \end{pmatrix} \begin{pmatrix} s \\ c \end{pmatrix}_\xi = \begin{pmatrix} q_w/\phi \\ 0 \end{pmatrix} \qquad (7)$$

with two constant states as initial data

$$\begin{pmatrix} s \\ c \end{pmatrix}_{(\xi,t=0)} = \begin{cases} (s_L, c_L)^T & , \quad \text{for } \xi < 0 \\ (s_R, c_R)^T & , \quad \text{for } \xi > 0 \end{cases} \qquad (8)$$

The Riemann problem (7,8) has a complex structure, which becomes apparent by considering the eigenvalues $\lambda^s = f_s$, $\lambda^c = f/(s+a_c)$ and eigenvectors $e^s = (1,0)^T$, $e^c = (f_c, \lambda^c - \lambda^s)^T$ of the Jacobian matrix associated with (7). The problem is *not strictly hyperbolic* (i.e., $\lambda^s(s,c)$ and $\lambda^c(s,c)$ may coincide for certain (s,c)) and *not genuinely nonlinear* (i.e., $e^s.\nabla\lambda^s$ and $e^c.\nabla\lambda^c$ may change sign). It is therefore not possible to apply the (standard) theory of hyperbolic systems (Lax, 1972).

However, Johansen and Winther (1988) have described the construction of a unique solution of the Riemann problem (7,8). The solution consists of constant states connected by smooth segments (*rarefaction waves*) and by discontinuities (*shock waves*). The rarefaction waves are associated with the integral curves of the eigenvectors e^s, e^c. Shock waves are (physically meaningful) discontinuities satisfying both the *Rankine-Hugoniot* jump condition and an *entropy* condition. The precise solution of the Riemann problem consists of a combination of (i) a c-shock and composite s-waves (if $c_L > c_R$) ; or (ii) c-rarefaction waves and composite s-waves (if $c_L < c_R$) A composite s-wave is a combination of an s-rarefaction wave and an s-shock that forms the solution of the (scalar) *Buckley-Leverett* problem $s_t + f(s,\bar{c})_\xi = 0$ for some fixed polymer concentration $\bar{c}$.

All shocks and waves arising from the solutions of different Riemann problems along the streamline are followed during the timestep of length Δt. Wave interactions are fully taken into account (Gmelig Meyling, 1990). Artificial smearing of discontinuities is virtually eliminated and timestep limitations of CFL-type (i.e., $\Delta t/\Delta\xi \; max(\lambda^s, \lambda^c) \le 1$) are avoided.

The implicit time-discretization $\frac{\partial s}{\partial t} \approx (s^{n+1} - s^n)/\Delta t$, $\frac{\partial (sc)}{\partial t} \approx ((sc)^{n+1} - (sc)^n)/\Delta t$ turns the *parabolic* equations (5) into a coupled system of elliptic equations. These equations are also conveniently discretized by mixed finite elements and solved by multigrid. The coupling between s and c and the nonlinearities in the diffusion tensors $\mathcal{D}^s(s,c)$, $\mathcal{D}^c(s,c)$ are handled by applying iteration.

Numerical Result

The figure illustrates a two-dimensional simulation of polymer flooding by the numerical method. Water (in gray) and polymer (in black) are injected at the lower-left corner of the square domain (a symmetry element of an oil reservoir). Oil (in white) is produced at the upper-right corner. The polymer slug is preceded and followed by a water flush. The reservoir is assumed to be homogeneous (i.e., constant permeability and porosity). *Local grid refinement* is used to increase the resolution of the oil/water and water/polymer interfaces. At the same time, the total number of grid blocks is reduced. Fine grid blocks (obtained by repeatedly subdividing coarser blocks) move along with the polymer slug and the water front as the displacement process continues.

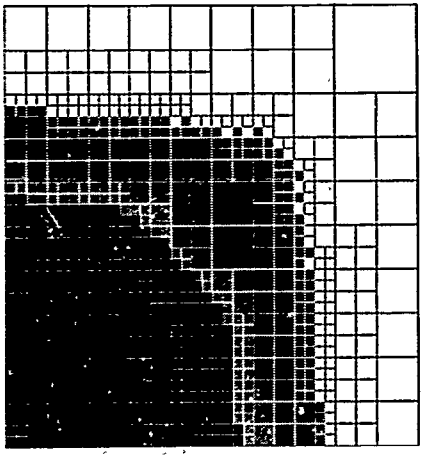


Fig. Numerical simulation of water preflush, polymer slug and water postflush for a homogeneous reservoir.

References

Gmelig Meyling, R.H.J., Numerical methods for solving the nonlinear hyperbolic equations of porous media flow, In: *Proceedings of the Third International Conference on Hyperbolic Problems*, Uppsala, 1990 (to appear).

Johansen, T. and Winther, R., The solution of the Riemann problem for a hyperbolic system of conservation laws modeling polymer flooding, SIAM J Math Anal 19 (1988), 541-566.

Lax, P.D., Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, *SIAM Regional Conference Series in Applied Mathematics*, Vol. 11, 1972.

Schmidt, G.H. and Jacobs, F.J., Adaptive local grid refinement and multigrid in numerical reservoir simulation, J Comput Phys. 77 (1988), 140-165.

COMPUTER SIMULATION OF A PHYSICO-MATHEMATICAL MODEL OF SOLAR ENERGY THERMAL STORAGE

V.L.ZIRA AND A. BUCUR
IEEE MEMBER No 1 142 561
Dept. of Physics, Polytechnical Institute of Bucharest,
Splaiul Independenței 313, Bucharest, Romania

Abstract. The knowledge of the time-distance evolution of a solar energy thermal storage system using rock beds crossed by an organic fluid allows the acquisition of important information for the optimum dimensioning of thermal solar energy tanks. Because of the difficulties arising during the analytical solution of the storage equations, we applied the numerical solution to the network method.

I. INTRODUCTION

The physical system considered is composed of a vertical cylinder containing the storage medium which is made up of particles of a very small diameter as compared to the diameter of the cylinder. The flow of the organic fluid is vertical with the warm part of the stock at the upper part of the cylinder and the cool part at the base in order to avoid the natural convection phenomena which lead to the destruction of the thermal stratification. The thermal transfer equations for physical system considered (storage equations) will be (Gicquel and others, 1979)

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} + w(t)\,\dfrac{\partial u(x,t)}{\partial x} + b_1(u-v) = 0 \\ \dfrac{\partial v(x,t)}{\partial t} + b_2(u-v) = 0 \end{cases} \qquad (1)$$

where $u = \dfrac{T_L - T_1}{T_2 - T_1}$; $v = \dfrac{T_S - T_1}{T_2 - T_1}$ are reduced temperatures; b_1 and b_2 are parameters characterizing the thermal properties of the fluid and the solid; $T_L(x,t)$ is temperature of the organic fluid; $T_S(x,t)$ is temperature of the solid medium; T_1 and T_2 are temperatures at the lower and upper ends of the stock, respectively; $w(t)$ is average flow velocity of the fluid.

The partial-derivative equations system obtained is a hyperbolic system. To solve it we use an improved version of the network method.

II. METHOD FOR THE NUMERICAL SOLUTION OF THE STORAGE EQUATIONS SYSTEM. COMPUTER. PROGRAM.

Because of the difficulties arising during the analytical solution of system (5), we applied the numerical solution with the network method (Ixaru, 1979; Absi and others, 1980).

The principle of the method consists in finding solutions along two families of characteristics

$dx = w\,dt$ for the fluid
$dx = 0$ for the solid

System (1) turns into systems (2) and (3)

$$\begin{cases} dx = w\,dt \\ du = b_1(v-u)dt \end{cases} \quad (2) \qquad \begin{cases} dx = 0 \\ dv = b_2(u-v)dt \end{cases} \quad (3)$$

We suppose the particular solutions already known

$$\begin{cases} u(t_0,x),\ v(t_0,x) \\ u(t,x_0),\ v(t,x_0) \end{cases}$$

In order to get easily processable results, the sets of data have been written in matrix form. This procedure was applied in order to specify the time-distance distribution of temperature in a condensed form. Thus, the variation of the temperature of the liquid is given in matrix form by $T_L(I,J)$, where the rows I indicate the behaviour at constant t=const. as a function of distance x and columns J the behaviour at constant distances x=const. as a function of time. This approach was extended for the functions $u(x,t)$ and $v(x,t)$, which give matrices $U(I,J)$ and $V(I,J)$.

The reduced temperatures u and v can thus be computed at the points where T_S and T_L are known:

$$u(i,j) = \frac{T_L(i,j)-T_1}{T_2-T_1};\ v(i,j) = \frac{T_S(i,j)-T_1}{T_2-T_1} \qquad (4)$$

A numerical solution by means of the network method requires sampling of input data. In order to get an acceptable accuracy for the solution, this sampling should be fine enough (16 x 16 points). The coefficients A_i (i= 1,6) have also been written in matrix form A(I,J) and have been used to compute the increments du and required by the network method.

The procedure is iterative. Starting from the given initial solution (u_0,v_0), an attempt is made to verify the system; if this does not fall within the required limits of accuracy, then $du = du(A_i)$ and $dv = dv(A_i)$ are computed and hence:

$$(u_1,v_1) = (u_0,v_0) + (du, dv)$$

is the new solution with which another attempt is made to verify the system.

The quantities du and dv are also written in matrix form: DU(I,J), DV(I,J).

The program computes the temperature distributions with double precision.

We note the following stages:
- computation of the fluid velocity w, of parameters b_1, b_2, of the initial U and V matrices using the initial T_S and T_L matrices;
- computation of the set of parameters A_i (i = 1,6);
- computation of the infinitesimal quantities dU(I,J) and dV(I,J);
- computation of matrices U(I,J) and V(I,J);
- computation of the final temperature distribution $T_{Sf}(I,J)$ and $T_{Lf}(I,J)$.

The results of the computations have been displayed in a table of values as two matrices U(I,J) and V(I,J) whose columns represent the time variation of the reduced temperatures at different distances from the origin (x = =const) and rows represent space variations of the reduced temperatures at different time moments (t = const).

To get a convenient solution of the system of partial derivative equations, we have passed from parameters $T_L(x,t)$, $T_S(x,t)$ to the functions $u(x,t)$, $v(x,t)$; these functions

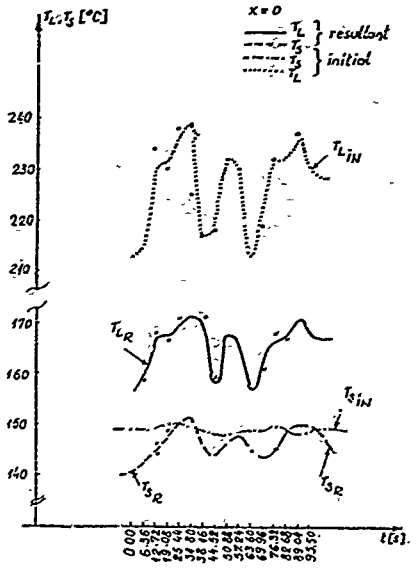


Fig.1. Dependence of temperatures T_L, T_S on time

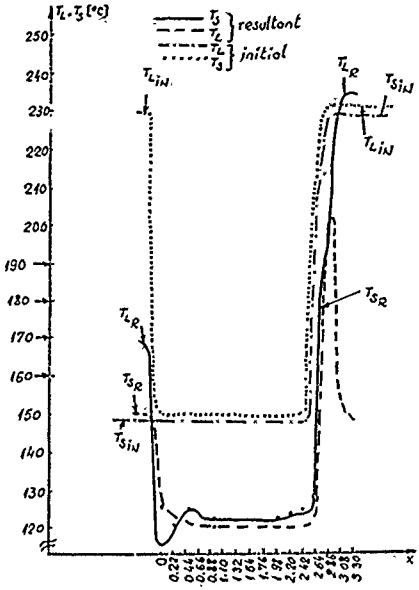


Fig.2. Dependence of temperatures T_L, T_S on distance

describe a ratio of the temperatures of the liquid and of the solid, respectively, to the difference in the temperatures of the upper and lower ends of the storage tank (T_2 and T_1, respectively).

We notice that $T_L(x,t)$ and $T_S(x,t)$ are warped surfaces in the plane $(x,t) \in [(x_0,x_f), (t_0,t_f)]$, corresponding to the different values of T_L, T_S at those point (Figs.1 and 2).

III. CONCLUSIONS

The main new points of this paper are:
- the use of a simple iteration method
- the use of matrix form.

The method used by the authors to obtain the temperature distribution has some important advantages. We notice the precision of the computations and the relatively high degree of generality which allows the applicability of the method to a variety of physical systems of interest for the thermal storage installations of the solar power plants. The comparison of these data with the experimental results will make possible the optimum dimensioning of thermal solar energy tanks.

REFERENCES

Gicquel, R., D. Harang and D. Schmell (1979) Rev.Gen.Therm., Fr., No 212-213, 527-535

Ixaru, L. (1979). Numerical Methods for Differential Equations and Applications. Roumanian Academy Publ. House, Bucharest

Absi, E., R. Glowinski, P.Lascaux and H. Veysseyre (1980). Méthodes Numériques dans les Sciences de l'Ingénieur. Dunod, Paris.

A SIMPLE NUMERICAL SCHEME TO DEAL WITH ONE-DIMENSIONAL ICE-FORMATION PROBLEM.

G. A. Vinnicombe and S. Misra
Department of Mechanical Engineering,
King's College London, London WC2R 2LS, U.K.

Abstract -A simple numerical scheme is described which deals with one-dimensional ice formation problem. The method is used here to numerically investigate how the rate of ice formation is affected by (a) convective heat transfer coefficient at the ice-water interface, (b) the effect of initial water temperature and (c) the effect of the coolant temperature. It is shown that the heat transfer coefficient and the initial water temperature have a limiting role to play in the ice formation process but, as expected, the effect of coolant temperature is very significant.

I. Nomenclature

t Temperature
t_w Water temperature
t_f Freezing point of water
t_c Coolant temperature
x Distance of ice-water interface from the plate
r Radial distance of ice-water interface from pipe
k_i Thermal conductivity of ice
k_w Thermal conductivity of water
f_o Heat transfer coefficient between coolant & plate
f_w Heat transfer coeff at the ice water interface
ρ_i Density of ice
ρ_w Density of water
c_i Specific heat capacity of ice
c_w Specific heat capacity of water
h_i Specific enthalpy of ice formation
K_i Thermal diffusivity of ice
θ Time
D Distance between the cooling surface and boundary of the water domain.

Subscripts
i Ice
w Water

II. INTRODUCTION

Ice-formation, an every day phenomenon, is easily achieved both naturally and under experimental conditions. However, the large amount of literature on the mathematical description of the problem, referred to as the Stefan Problem, is considerable and testifies to the fact that the proper prediction of the process is, on the other hand, not so easy. This literature is well reviewed by Crank [1], Salcudean and Abdullah [2] and Lunardini [3]. Because of its non-dependence on the phase change front the enthalpy method [4] is by far the most widely used method to deal with phase change problem. There are, though, two apparent difficulties in employing the enthalpy method for the ice formation problem. The first is the specific heat capacity discontinuity at the freezing temperature and the second is in accounting for the convective heat transfer at the ice-water interface. Both these problems have been successfully tackled [4,5] but not before indulging in tedious mathematical formulations.

In another method, the front tracking method, the position of the ice front is obtained at the end of each iteration. The difficulty with front tracking method, apart from its inherent weakness of being dependant on the position of the interface, is its extreme complexity in solving higher dimensional problems. Fortunately, though, for most engineering ice-formation applications, the one-dimensional result is sufficiently accurate for all practical purposes [5]. The advantage of this method is that the position of the ice front, which is the most important quantity, is obtained directly from the analysis and it is not necessary to infer it from an interpretation of the temperature profiles as is necessary with the enthalpy method.

The purpose of the present paper is to present a simple alternative front tracking method and to demonstrate its use by evaluating the role of convective heat transfer coefficient at the ice-water interface and the effect of initial water temperature and the coolant temperature on the rate of ice-formation. This method has the advantage of computing the position of the ice front directly but is much easier to apply than the front tracking method.

III FORMULATION

Consider a simple model of ice forming on a semi-infinite plate which is adjacent to coolant held at a constant temperature below the freezing temperature of water (Figure 1).

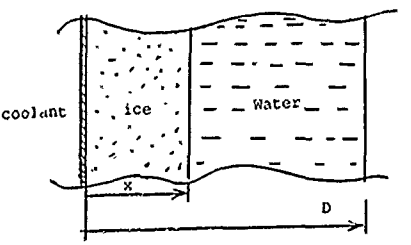


Figure 1

The water temperature is assumed to be uniform at t_w. The enthalpy balance equation at the ice-water interface yields the velocity of the ice front as:

$$\frac{\partial x}{\partial \theta} = \left[k_i \frac{\partial t}{\partial x}\Big|_x - f_w (t_w - t_f) \right] \Big/ \rho_i h_i \qquad (1)$$

Where $\frac{\partial t}{\partial x}\Big|_x$ is the local temperature gradient within the ice mass at the interface.

Now consider an element at the interface on the water side, in time $\partial\theta$ an enthalpy balance yeilds the rate of change of water temperature with time as :

$$\frac{\partial t_w}{\partial \theta} = \frac{- f_w \cdot (t_w - t_f)}{(D - x) \cdot \rho_w \cdot c_w} \qquad (2)$$

Equations (1) and (2) are a pair of first order simultaneous differential equations which can be solved to give the position of ice front and the water temperature during the freezing or thawing process.

For freezing around a pipe the corresponding equations are :

$$\frac{\partial r}{\partial \theta} = \left[k_i \frac{\partial t}{\partial r}\Big| - f_w (t_w - t_f) \right] \Big/ \rho_i h_i \qquad (3)$$

and

$$\frac{\partial t_w}{\partial \theta} = \frac{- f_w \cdot (t_w - t_f)}{(D - r) \cdot \rho_w \cdot c_w} \qquad (4)$$

IV SOLUTION

Equations (1) and (2) can be solved numerically using any of the standard procedures for initial value problem. The only problem is the term $k_i \cdot \partial t/\partial x$ in equation(1). At the start, when there is no ice then :

$$k_i \cdot \frac{\partial t}{\partial x}\Big|_0 = f_w \cdot (t_f - t_c)$$

Thereafter the temperature gradient can be obtained from the solution of the conduction equation in the ice . ie.

$$\frac{\partial t}{\partial \theta} = K_i \cdot \frac{\partial^2 t}{\partial x^2} \qquad (5)$$

This can be solved, most conveniently using a finite difference equivalent of (5) with the boundary conditions :

$$x = 0 \quad ; \quad t = t_c$$

$$x = X \quad ; \quad t = t_f$$

Thus the solution is a combination of, for example, the Runge Kutta Method and the finite difference method.

A typical result is shown in figure 2.

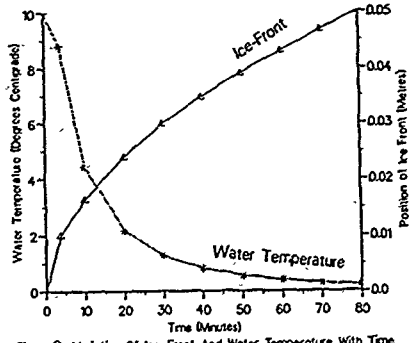


Figure 2: Variation Of Ice-Front And Water Temperature With Time

V TEST PROBLEM

The stated formulation was tested by simulating the following conditions: Ice is formed on a flat vertical plate adjacent to coolant which is maintained at a constant temperature. The plate is in an insulated water tank and the water is initially at a uniform temperature. The dimensions of the plate and the tank are so chosen and the conditions are such that the assumption of one-dimensional freezing is justified. The various parameters and conditions used for computations of the results are included in table 1.

TABLE I
PARAMETERS FOR THE TEST PROBLEM

Heat transfer coefficient at the coolant/ice boundary (f_o)	2.0 kW/m^2/K
Thermal Conductivity of Ice (k_i)	2.2$\cdot 10^{-3}$kW/mK
Density of ice (ρ_i)	917 kg/m^3
Specific heat capacity of ice (c_i)	2.0 kJ/kgK
Sp. enthalpy of ice formation (h_i)	-333.5kJ/kg
Thermal conductivity of Water (k_w)	.6$\cdot 10^{-3}$kW/mK
Density of Water (ρ_w)	1000 kg/ m^3
Sp. heat capacity of water (c_w)	4.18kJ/kgK
Freezing point of water (t_f)	0°C
Distance between plate and boundary (D)	1.0m.

To assess the role of convective heat transfer coefficient at the ice water interface (f_w), the water temperature (t_w) was initially 10°C and the coolant temperature (t_c) kept at -40°C. The various values of f_w used ie 0.01 kW/m^2/K to 4.0 kW/m^2/K were chosen to cover the entire spectrum of likely practical values of f_w. The ice front position at different times for the various values of f_w is shown in table 2. Because the water temperature was also changing there was no maximum ice thickness and ultimately, all the water will become ice.

TABLE II
POSITION OF ICE FRONT AT DIFFERENT INTERVALS WITH VARYING f_w

TIME (MIN.)	POSITION OF ICE FRONT (MM)					
f_w	10.00	17.08	16.33	15.42	15.25	15.53	16.02
	19.99	24.73	23.86	23.70	23.63	23.82	24.15
	29.95	30.56	29.72	29.88	29.82	29.97	30.24
	40.00	35.47	34.69	35.02	34.97	35.10	35.33
	50.07	39.76	39.10	39.51	39.47	39.58	39.78
	60.14	43.65	43.09	43.56	43.52	43.62	43.80
	70.21	47.21	46.77	47.26	47.22	47.31	47.49
f_w	>	0.01	0.1	0.5	1.0	2.0	4.0

To evaluate the effect of initial water temperature, the coolant temperature was put at -40°C, f_w at 0.1 kW/m²/ K and the water temperature varied from 5.0°C to 40°C. The result is shown in table 3.

TABLE III
POSITION OF ICE FRONT AT DIFFERENT INTERVALS AND VARYING WATER TEMPERATURE

TIME (MIN.)	POSITION OF ICE FRONT (MM.)					
f_w	10.00	16.77	16.33	15.90	15.47	14.59	13.70
	19.99	24.41	23.86	23.32	22.80	21.77	20.75
	29.95	30.30	29.72	29.15	28.61	27.56	26.56
	40.00	35.28	34.69	34.12	33.58	32.56	31.60
	50.07	39.69	39.10	38.53	38.01	37.02	36.10
	60.14	43.66	43.09	42.54	42.04	41.10	40.22
	70.21	47.33	46.77	46.23	45.75	44.83	44.01
t_w (°C)	>5.0		10.0	15.0	20.0	30.0	40.0

In a similar way the effect of coolant temperature was evaluated by putting f_w = 0.1 kW/m²/K, t_w = 10° C and varying the coolant temperature from -10°C to -40°C. The result is given in table 4.

TABLE IV
POSITION OF ICE FRONT AT DIFFERENT INTERVALS AND VARYING COOLANT TEMPERATURE

TIME (MIN.)	POSITION OF ICE FRONT (MM.)				
f_w	10.00	6.99	9.06	10.63	13.70	16.33
	19.99	9.63	12.93	15.76	20.16	23.86
	29.95	11.91	16.56	19.95	25.25	29.72
	40.00	14.55	19.74	23.55	29.59	34.69
	50.07	16.97	22.56	26.75	33.44	39.10
	60.14	19.16	25.11	29.64	36.92	43.09
	70.21	21.16	27.44	32.32	40.13	46.77
t_c (°C)		-10.00	-15.00	-20.00	-30.00	-40.00

VI RESULTS AND DISCUSSION

A study of the results in table 2 shows that for all practical purposes the value of f_w has no effect on the rate of ice formation. A lower value of f_w results in initial higher rate of ice formation but as the ice builds up, the rate decreases compared to higher values of f_w. The overall rate of ice formation is marginally more in the case of higher f_w values but for all practical purposes the difference is insignificant. Results from table 3 show, as would be expected, that with increasing water temperature the rate of ice formation decreases. However it is noted that the effect is quite small.

Looking at these results together, it can be concluded that heat transfer coefficient at the ice-water interface and the initial temperature of water has little effect on the overall rate of ice formation. It can be concluded that, for all practical conditions, of the heat extracted via the ice, the portion required to cool the water is small and can be neglected.

Results from table 4 shows that the coolant temperature has a significant effect on the rate of ice formation.

VII CONCLUSION

The method reported in the paper is quite simple and easy to use yielding a mathematical insight into the ice formation problem. Data runs done with the method gave useful information on the role of the different parameters and input conditions on the rate of ice formation. The value of convective heat transfer coefficient at the ice-water interface has a very limiting role to play in the process and so too the initial water temperature. The coolant temperature is the most important parameter in the ice formation problem.

VIII REFERENCES

1. J.Crank, How to deal with moving boundaries in thermal problems. In Numerical Methods in Heat Transfer. (Edited by R.W. Lewis, K. Morgan and O. C. Zienkiewicz pp. 177-200, Wiley, New York, 1981.

2. M.Salcudean and Z.Abdullah, On the numerical modelling of heat transfer during solidification processes. Int.J. numer.methods eng., vol. 25, 445-473, 1988.

3. V. Lunardini, Heat Transfer in Cold Climates, Van Nostrand Reinhold, New York, 1981.

4. V. Voller and M. Cross, Accurate solutions of moving boundary problems using the enthalpy method. Int. J. Heat Mass Transfer, Vol. 24, pp. 545-556, 1981.

5. T. Hirata, R.R.Gilpin, K.C. Cheng and E.M. Gates, The steady state ice layer profile on a constant temperature plate in a forced convection flow -I. Laminar Regime. Int. J. Heat Mass Transfer, Vol. 22, pp. 1425-1433, 1979.

Acknowledgement : One of the authors (S.Misra) is a holder of the Edmund Davis Scholarship of the University of London and the ORS award.

MIXED FINITE ELEMENT AND NEURAL NETWORK METHODS OF VISUAL RECONSTRUCTION

D. Suter
Department of Computer Science and Computer Engineering
La Trobe University
Bundoora 3083
Australia

Abstract: The regularization formulation of visual reconstruction problems is approximated using a mixed finite element approach. This lends itself to interesting solution methods: digital and analog. In particular, the analog networks are similar to but more simple than those proposed previously.

I. INTRODUCTION

Visual Reconstruction is concerned with the recovery of information from images and includes approaches such as the recovery of depth from stereopsis or the recovery of shape from shading. A common paradigm is to formulate such problems, within the framework of regularization theory, as inverse problems. Furthermore, using an analogy between energy in an electrical circuit and the objective functional of the regularization formulation, many analog networks have been derived and promoted as neural networks for machine vision.

This paper uses a mixed finite element approximation to the regularized problem. From a purely computational point of view, this has several advantages. Firstly, lower order finite element basis functions can be used. Secondly, such an approach allows various derivatives to be explicitly reconstructed at the same time as the function of interest. These estimates of the derivatives of the function can be useful for segmentation or in the search for features. The approach can also be shown to naturally incorporate aspects found to be useful from purely experimental studies [1].

Another major contribution of this work is to investigate analog network schemes for solving the the mixed finite element formulation. The setting we propose already generalizes that of Harris [2]. We have used variants of the Arrow/Hurwicz approach to solving saddle point problems (first proposed as a neural network model by Platt [3]) to derive a whole variety of analog network structures.

We conclude by discussing work in progress to find more efficient digital approaches to solving our mixed finite element formulations.

II. VISUAL RECONSTRUCTION AND FINITE ELEMENTS

Suppose we are interested in recovering a quantity represented by the function ψ and, from our model of the image formation process, we can write down the relationship between this quantity and the observed data d as $A(\psi, d) = 0$. Typically, even if one can solve this system, the inverse nature of the problem makes the solution unstable and we generally seek a regularized solution that minimizes:

$$J = \int_\Omega (A(\psi, d))^2 + (D(\psi))^2 \, dx \, dy \qquad (1)$$

where the last term is a penalty term that encourages smoothness (and D is an operator that usually is composed of various derivative operators). Much of the study of these approaches in vision has concentrated upon the recovery of surfaces from sparse depth estimates. This model problem is the most simple of the visual reconstruction problems and reduces to a generalized spline fitting problem. For example, a commonly studied instance is the problem of minimizing:

$$J = \frac{\beta}{2} \sum_{i \in C} (\psi_i - d_i)^2 + \frac{1}{2} \int_\Omega \frac{\partial^2 \psi}{\partial \psi^2} + 2 \frac{\partial^2 \psi}{\partial x \partial y} + \frac{\partial^2 \psi}{\partial y^2} \, dx \, dy \qquad (2)$$

where β is a constant related to the reliability of the data, C is a set of discrete points at which we have constraints upon the solution (e.g. depth values in stereo) and the integral is a second order smoothness (Sobolev) seminorm.

Even though fast algorithms exist for spline fitting (particularly in one dimension), in machine vision the research has concentrated on solving discrete approximations through local iterative approaches. A typical approach is to discretize using finite elements and then solve using multigrid iterative approaches [4]. Such an emphasis reflects the fact that we are interested in this problem as a model problem only (the particular problems of direct interest usually involve more complex and possibly nonlinear operators A). Furthermore, local iterative methods are well suited to parallel implementation and are neural network-like. Indeed, even more exotic analog network methods have been devised to even more closely resemble biological approaches to vision [5] and, at the same time, improve speed and robustness.

III. MIXED FINITE ELEMENTS

It is easy to see that the Euler-Lagrange equation for 2 is closely related to the biharmonic operator (fourth order), and hence quite complex finite elements are required [4]. The author has used the mixed finite element formalism to reduce the order of the differential operators appearing in the smoothness functional [6]. Consider the one dimensional example[1] of minimizing:

$$J = \frac{\beta}{2}\sum_{i\in C}(\psi_i - d_i)^2 + \frac{1}{2}\int_\Omega \left(\frac{d\psi}{dx}\right)^2 + \left(\frac{d^2\psi}{dx^2}\right)^2 dx \quad (3)$$

The start of our approach is to consider the derivatives as themselves being unknown functions that we seek to reconstruct. Thus, we seek to minimize:

$$J = \frac{\beta}{2}\sum_{i\in C}(\psi_i - d_i)^2 + \frac{1}{2}\int_\Omega u^2 + p^2\, dx \quad (4)$$

subject to $u = \frac{d\psi}{dx}$, and $p = \frac{du}{dx}$. There are two main ways in which to convert the above constrained optimization problem into an unconstrained problem One method is to use penalty terms (and this leads directly to the formulation of Harris [2], for which I have derived an analog network structure), and the other is to use Lagrange Multiplier approaches. However, it is also possible to combine the two approaches in an Augmented Lagrangian formulation [7]. For our problem, this would lead to the problem of seeking extrema of the functional:

$$J = \frac{\beta}{2}\sum_{i\in C}(\psi_i - d_i)^2 + \int_\Omega \tfrac{1}{2}(u^2 + p^2) + \lambda_u\left(\frac{d\psi}{dx} - u\right) + \lambda_p\left(\frac{du}{dx} - p\right) + \frac{\rho}{2}\left(\left(\frac{d\psi}{dx} - u\right)^2 + \left(\frac{du}{dx} - p\right)^2\right) dx \quad (5)$$

where ρ is a positive constant, and λ_u, λ_p are Lagrange Multipliers. It would appear that we pay the price for low order elements in terms of an increase in the number of nodal points and in terms of lower order accuracy in the reconstruction. In addition, it seems necessary to calculate the Lagrange Multipliers. However, it is easy to show that $\lambda_u = u$ and $\lambda_p = p$. Furthermore, it may be possible to use post-processing [8] to improve the accuracy of approximation.

If $\rho = 0$, the above problem reduces to a Lagrangian formulation (and the solution is generally a saddle point rather than a minima). It is possible to show that if ρ is large enough, the problem essentially becomes a penalty based formulation (and the solution is a minima). For simplicity we will set $\rho = 0$ in the sequel.

By eliminating the multipliers, and reformulating the constraints into weak constraints, it is possible to reformulate as the extremization of:

$$L = \frac{\beta}{2}\sum_{i\in C}(\psi_i - d_i)^2 + \int_\Omega \tfrac{1}{2}(u^2 + p^2) + u\frac{d\psi}{dx} + \frac{dp}{dx}\frac{du}{dx}\, dx \quad (6)$$

We generally discretize such formulations using linear elements to arrive at the mixed finite element formulation in terms of the nodal variables ψ_i, u_i and p_i..

IV. ANALOG NETWORKS

Platt [3] and Szyman [9] independently proposed a variant of the Arrow/Hurwicz approach [10] for finding saddle points of a functional. In essence, this is a form of gradient descent on the primary variables and ascent on the secondary variables. In our case, this corresponds to the scheme:

$$\frac{d\psi_i}{dt} = -\frac{\partial L}{\partial \psi_i},\ \frac{du_i}{dt} = +\frac{\partial L}{\partial u_i},\ \frac{dp_i}{dt} = +\frac{\partial L}{\partial p_i} \quad (7)$$

We have integrated such equations for the problem 6 and the results are shown in figure 1.

It is relatively straightforward to map this analog approach onto hardware using transconductance amplifiers [11]. Simulations of these circuits using SPICE show that the circuits correctly operate. These circuits are considerably simpler than those previously suggested [2].

V. DIGITAL IMPLEMENTATION

The analog methods in the previous section, when simulated on a digital machine, require many iterations and

[1] Although, for machine vision, we are clearly interested in formulations over a two dimensional domain Ω, we will restrict ourselves to one dimensional examples in much of the sequel The generalization to higher dimensions is straightforward but more cumbersome.

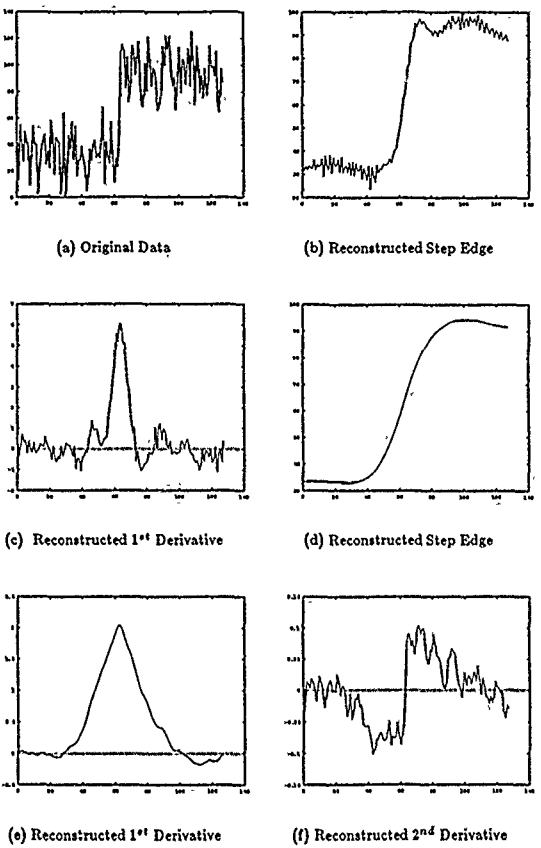


Figure 1: Reconstructed Step Edge
The original step edge (a) is reconstructed showing. (b) the function, (c) first derivative, after 1000 iterations, and (d) the function, (e) the derivative, and (f) the second derivative after 10000 iterations. Note the clear peak in the first derivative and the zero crossing in the second.

are thus relatively slow. In this section we consider digital methods.

We consider now, for simplicity, the discrete form of one dimensional regularization using only the first order smoothness term (discarding the second order smoothness term from 6): one obtains a set of Euler-Lagrange equations that can be written in matrix form:

$$\begin{pmatrix} \beta I_1 & D^T \\ D & -I_2 \end{pmatrix} \begin{pmatrix} \psi \\ u \end{pmatrix} = \begin{pmatrix} \beta d \\ 0 \end{pmatrix} \qquad (8)$$

where D is a discrete first derivative operator, and I_1 and I_2 are similar to the identity matrices. For the case $\rho \neq 0$ this becomes

$$\begin{pmatrix} \beta I_1 + \rho DTD & (1-\rho)D^T \\ (1-\rho)D & -(1-\rho)I_2 \end{pmatrix} \begin{pmatrix} \psi \\ u \end{pmatrix} = \begin{pmatrix} \beta d \\ 0 \end{pmatrix} \qquad (9)$$

We are currently studying methods of solving these by using conjugate direction (including hierarchical basis functions [12] or other methods for pre-conditioning [13]) and evaluating with alternative ways of dealing with the indefinite nature of the system [14].

VI. CONCLUSION

There are two major contributions in this work so far. One is the introduction of mixed finite element methods into machine vision. The other is in the investigation of novel analog solution methods for the mixed formulations.

These developments have inspired current work on efficient (parallel) solution methods of the related indefinite systems using digital hardware.

References

[1] B. K. P. Horn, "Height and gradient from shading," AI Memo 1105, MIT, May 1989.

[2] J. G. Harris, "A new approach to surface reconstruction: the coupled depth/slope model," in *Proc. First Int'l Conference on Computer Vision, London*, pp. 277–283, 1987.

[3] J. C. Platt and A. H. Barr, "Constrained differential optimization," in *Proceedings of the 1987 Neural Information Processing Systems Conference, Denver Colorado*, (New York), pp. 612–621, AIP, Nov. 1987.

[4] D. Terzopoulos, "Multi-level reconstruction of visual surfaces: variational principles and finite element representations," AI Memo 671, MIT, April 1982.

[5] C. Mead, *Analog VLSI and Neural Systems*. Reading MA: Addison-Wesley, 1989.

[6] D. Suter, *Co-operative Algorithms in Machine Vision: Models, Problem Formulation, and Neural Network Implementations*. PhD thesis, Dept. Comp. Sci. and Comp. Eng., Bundoora, 3083, Australia, August 1990.

[7] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Amsterdam: North-Holland, 1983.

[8] J. H. Bramble and J. Xu, "A local post-processing technique for improving the accuracy in mixed finite-element approximations," *SIAM J. Numerical Analysis*, vol. 26, pp. 1267–1275, Dec. 1989.

[9] J. A. Snyman, "A parameter-free multiplier method for constrained minimization problems," *Journal of Computational and Applied Mathematics*, vol. 23, pp. 155–168, 1988.

[10] P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization. Cambridge Texts in Applied Mathematics*, Cambridge: Cambridge University Press, 1988.

[11] D. Mansor, "An analog circuit for first order regularization," La Trobe University technical report 2/91, March 1991.

[12] H. Yserentant, "Two preconditioners based on the multi-level splitting of finite element spaces," *Numer. Math.*, vol. 58, pp. 163–184, 1990.

[13] J. H. Bramble and J. E. Pasciak, "A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems," *Mathematics of Computation*, vol. 50, pp. 1–17, January 1988.

[14] O. Axelsson, "Preconditioning of indefinite problems by regularization," *SIAM J. Numer. Analysis*, vol. 16, pp. 58–59, February 1979.

APPLICATION OF POWER SYSTEMS ANALYSIS PACKAGE (PAP) TO EREĞLI IRON AND STEEL WORKS, INC., EREĞLI

Osman SEVAIOĞLU,

Electrical and Electronics Engineering Department,
Middle East Technical University,
Ankara, Turkey

Orhan YILDIRIM

Ereğli Iron and Steel Works, Inc.,
Kdz. Ereğli,
Zonguldak, Turkey

ABSTRACT

In this paper, application of Power System Analysis Package (PAP) to the distribution system of Ereğli Iron and Steel Works, Inc. is described.

The major aim of the paper is to discuss the modifications made on the solution algorithms in the Power Systems Analysis Package, PAP, in order to adapt them to an industrial distribution system.

The paper also includes a short description of the organization of new cable data base, known as the low-voltage cable library. In the last part of the paper, a system plot routine developed during the research is described.

1. POWER SYSTEMS ANALYSIS PACKAGE: PAP

1.1. Development

Initial research on the Power Systems Analysis Package (PAP) was started at the University of Manchester, Institute of Science and Technology (UMIST) in 1973. During this research, two algorithms were developed for the steady-state solution of electric power systems.

The first of these algorithms is for the optimum power flow problem, which deals mainly with the constrained minimization of a the system operating cost /7/. The algorithm is based the well-known method of Dommel-Tinney /6/ and has proved quite successful for the solution of systems operating under stress as well as under normal steady-state conditions.

The other algorithm is the Fast-Decoupled Load-Flow, the highly efficient, fast, and reliable solution algorithm for the solution of the power systems /4/.

After 1972, a parallel research was started in the Middle East Technical University. The main outcome of this research was the Balanced/Unbalanced Short-Circuit Analysis and the power system data base IDSPS /11/.

Another integral part of PAP is the transient and dynamic stability analysis routine developed in 1973 /12/.

The first industrial application of PAP was realized in the Planning and Coordination Division (PKD) of the Turkish Electricity Authority (TEK) in 1979. The package is still in use for the routine system analysis tasks in PKD.

The package was then adapted to Çukurova Electricity Utility for system analysis and planning studies on the southern regions of Turkey, and is being used successfully.

1.2. Fast-Decoupled Load-Flow Algorithm

The Fast-Decoupled Load-Flow equations are obtained by solving the well-known polar power mismatch equations using Newton's Method /4/. The method is a straightforward application of the general Newton solution algorithm for solving the sparse real equation

$$\begin{bmatrix} H & N \\ J & L \end{bmatrix} \begin{bmatrix} \Delta\theta \\ \Delta V/V \end{bmatrix} = \begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} \qquad (1)$$

where, the submatrices in the coefficient matrix consist of the jacobians of active and reactive power mismatches with respect to angles and scaled voltage magnitudes, respectively.

Several approximations can be made in the above solution algorithm /4/. These approximations may be briefly listed as follows:

(i) neglect the off-diagonal submatrices N and J by applying the decoupling principle, hence leaving two separated equations.

(ii) keep the jacobian submatrices H and L constant at the initial solution point, where they are evaluated and factorized only once,

(iii) approximate the cosine and sine functions to be evaluated during the calculation of the jacobian submatrices H and L by unity and radians of the angles, respectively,

With these approximations the matrices H and L become strictly the elements of the nodal admittance matrix B. These matrices are then called B' and B'', respectively. In terms of these matrices further approximations are made as follows:

(iv) omit from B' the shunt reactances, and off-nominal transformer tap adjustments,

(v) transfer the diagonal left-hand side voltage matrices which arise from the division of measurements /4/, to the right-hand sides as inverses,

(vi) neglect the remaining diagonal voltage matrices on the left-hand sides of the equations by assuming that the voltages in a system operating under normal operating conditions are near unity,

(vii) neglect the series branch resistances while calculating the elements of B'.

The resulting Fast-Decoupled Load-Flow equations then become

$$B' \Delta\theta = \Delta P/V \quad (2)$$

$$B'' \Delta V/V = \Delta Q/V \quad (3)$$

1.3. Limitations of the Fast-Decoupled Load-Flow Algorithm

Success of the fast-decoupled Load-Flow algorithm, is strongly based on two features of the system: the X/R ratio, and the operating conditions. The X/R ratio must be high, and the system must be in its normal steady-state operating condition. Violation of any or both of these conditions may easily lead the algorithm to a convergence difficulty or even to a divergence. Performance of the algorithm is significantly degraded as this ratio is reduced. In extreme cases where this ratio falls below 0.1, divergence is generally observed.

Table 1 lists the performances of the conventional Newton Load-Flow algorithm and the Fast-Decoupled algorithm for the Eregli distribution system, where the X/R ratio is progressively reduced from its nominal value to 10% by multiplying branch reactances with a varying scale factor. The tests are carried out without any control on transformer taps and generator reactive power outputs. Convergence tolerance used in the tests is 0.001 MW/MVAR for both tests. Tests on the conventional Newton algorithm are carried out by using the optimum load-flow module in PAP.

Table 1. Number of Iterations for Eregli System with Progressively Reduced X/R Ratio.

X-Scale Factor	Newton-Raphson Load-Flow	Fast-Decoupled Load-Flow
1.000	5	7
0.500	5	12
0.250	5	35
0.200	5	div.
0.166	5	div.
0.125	6	div.
0.111	6	div.
0.100	6	div.
0.080	7	div.

It may be concluded from the above test that convergence of the Fast-decoupled Algorithm is unreliable for low voltage distribution systems. The algorithm exhibits severe convergence difficulties as the X/R ratio is reduced. hence cannot be used for the analysis of Eregli System in its original form.

Similar convergence difficulties are observed when branch resistances are multiplied by a scale factor which is progressively increased from unity to 5.0. The situation is even worse in this case, since branch impedances are increased, thus increasing bus angles.

1.4. The BX Scheme

To eliminate the limitations of the Fast-decoupled load-flow algorithm an alternative version is proposed in Ref /5/. This version, also called the BX Scheme, provides a much better convergence performance while retaining all the attractive features of the Fast Decoupled Algorithm.

The main principle behind this alternative version is surprisingly simple: neglect the resistances of the series branches while calculating the jacobian matrix B'', rather than B'.

The BX Scheme is implemented to the Fast-Decoupled Load-Flow Algorithm in PAP, and its performance is tested. Similar to the test carried out for the Fast-Decoupled Load-Flow, the X/R ratio is progressively reduced from its nominal value to 10% by multiplying branch reactances with a varying scale factor. The remaining conditions in the tests are not varied so to compare the performance of the two algorithms. In order to be exhaustive, the operating condition of the system is varied in order to observe its effect on the performance.

Performance indices obtained by using the BX scheme is listed in Table 2. As may be seen from the Table, a highly improved performance is obtained as reported in the literature /5,10/.

Table 2. Number of Iterations for the Eregli System with Progressively Reduced X/R Ratio using the BX Scheme.

X-Scale Factor	Fast-Decoupled Load-Flow using the BX Scheme.
1.000	7
0.500	7
0.250	9-10
0.200	9-11
0.166	10-11
0.125	13-19
0.111	11-13
0.100	13-15
0.080	15-19

1.5. Short Circuit Analysis

The short circuit analysis module in PAP is used to calculate post-fault bus voltages, line, cable, transformer, generator, and induction motor currents, and interruption duties of circuit breakers under balanced or unbalanced fault conditions.

Most of the features of the program are quite similar to that of a conventional short circuit analysis program, and hence are not described here. Instead, the modifications made on the program for adapting it to the Eregli system will be described.

a. Momentary Duties

The momentary duty of a circuit breaker is the maximum transient current expected to pass within the fault clearing period. Depending on the system X/R ratio and the time instant at which the fault is initiated, this current may be as high as 1.7 times the normal fault current.

In general, for the sake of simplicity, only the most severe fault condition: a three-phase-to-ground type fault is considered.

Momentary duty characteristics are empirical curves, hence they are represented in the algorithm by a fourth-order polynomial.

b. Interruption Duties

The interruption duty of a protective equipment is the maximum fault current at the instant of clearing. This current depends on a ratio, known as No AC Decrement Ratio (NACD), as well as its clearing time and the system X/R ratio. This ratio is defined as:

$$NACD = I_{ext} / I_{fault} \qquad (4)$$

where, I_{ext} is the total fault current received from external sources, that is from the utility system,

I_{fault} is the total symmetrical fault current flowing to ground at the fault point.

Multiplying factor curves for interruption duties of 4-cycle, 3-cycle, and 2-cycle circuit breakers are represented by again fourth-order polynomials.

2. Low-Voltage Cable Library

The Power System Analysis Package (PAP) has an extensive data base which stores all relevant physical and electrical data concerning the transmission and generation system. A detailed description of the organization of this data base is given in /9,11/. This section therefore, is devoted only to the description of the low-voltage cable library developed for the Eregli distribution system.

The low-voltage cable library is so designed that a user can easily access the electrical characteristics of a low-voltage cable or line used in his study. To perform a load-flow or a short-circuit analysis, the user needs to specify only the length and type of the cables.

Table 3 lists the part of the low-voltage cable library corresponding to 2.4. kV cables.

Table 3. Low-Voltage Cable Library (6.3 kV)

Cable Size mm-sq	Resis. Ohm/Km	React. Ohm/Km	Susc. Mho/Km	(0)Res. Ohm/Km	(0)Reac. Ohm/Km	Capacity Amp.
3*6	2.9199	.1640	.0000	8.7598	.4888	
3*10	1.9652	.1673	.0000	5.8957	.5052	
3*16	1.5059	.1673	.0000	4.4948	.4987	
3*25	.9580	.1575	.0000	3.8287	.6332	125.0
3*35	.6594	.1509	.0000	3.2808	.7513	150.0
3*40	.5873	.1476	.0000	2.9298	.7415	158.3
3*50	.4790	.1476	.0000	2.4016	.7349	175.0
3*63	.3871	.1476	.0000	1.9291	.7448	204.2
3*70	.3510	.1476	.0000	2.8150	1.1713	220.0
3*80	.3051	.1444	.0000	2.4409	1.1549	246.6
3*95	.2608	.1419	.0000	2.0915	1.1327	260.0
3*100	.2461	.1411	.0000	1.9751	1.1253	267.0
3*120	.2067	.1312	.0000	1.6535	1.1024	295.0
3*125	.2001	.1378	.0000	1.5945	1.0958	300.7
3*150	.1706	.1378	.0000	1.3550	1.0991	329.6
3*160	.1575	.1378	.0000	1.2566	1.1024	341.2
3*185	.1444	.1378	.0000	1.2500	1.2300	370.0
3*200	.1280	.1345	.0000	1.2730	1.3583	387.3
3*250	.1050	.1345	.0000	1.0400	1.3287	440.0

3. SYSTEM PLOT ALGORITHM

Results obtained from the load-flow and short-circuit analyses are plotted by a plot routine integrated into PAP. This routine is designed to plot the overall single line diagram of the Eregli System which is of about 160 X 80 cm dimension, plotted by a large size plotter.

Output of this routine includes the power flows in cables under steady state operating conditions, generator loadings, transformers with tap adjustments.

Plot output obtained from the short-circuit study is exactly of the same form as that for the load-flow, except that flows and voltages in this case corresponds to fault currents, post fault voltages, respectively.

4. EREGLI IRON AND STEEL WORKS DISTRIBUTION SYSTEM

Eregli Iron and Steel Works Distribution System is used to supply the power demand of the overall complex distributed in a 3.5 square-km area. The peak power demand of the site is about 125 MW, of which 75 MW is received from the 154 kV utility buses. The remaining power is generated by three thermal generating units installed at the site. The peak power demand is expected to reach 200 MW level in 1995 when the expansion plan is completed.

The system configuration is mainly radial, although some links are drawn as security lines which are kept open under normal circumstances. Bus, branch and transformer totals in the system are about 170, 68, and 120, respectively. The nominal voltage levels in the system are 13.8, 2.4, 0.4 kVolts. The system is supplied from the utility buses by four tie lines, two of them being at 66 kV and two at 154 kV. series reactors are installed mainly for short-circuit current limitation purposes.

The utility system beyond the tie line buses is represented by Thevenin generators, of which the subtransient reactances are calculated using the bus short-circuit levels.

5. CONCLUSIONS

The following conclusions may be drawn from the experience gained in applying PAP to Eregli System.

(i) Fast-Decoupled Load-Flow Algorithm is unreliable for the solution of distribution systems with low X/R ratio.

(ii) The BX Scheme performs much better, hence is more suitable to distribution systems.

(iii) A distribution system engineer always tends to deal with the real, physical aspects of problems, hence tries to avoid himself to be in a sophisticated computational environment with a large volume of data. Therefore, facilities such as low-voltage cable library are highly useful.

(iv) Due to size of the system, plotting is an essential requirement in distribution system analysis.

ACKNOWLEDGEMENT

This research reported in this paper was sponsored by the Eregli Iron and Steel Works, Inc., Eregli. The authors express their appreciations for the support.

REFERENCES

/1/ "Interrupting Capability Factors for Reclosing Services for AC High-Voltage Circuit Breakers", ANSI Standard C37.07.

/2/ "Methods for Determining the RMS value of a Sinusoidal Current Wave and a Normal-frequency Recovery Voltage and for Simplified Calculation of Fault Currents", ANSI Standard C37.5.

/3/ "Calculation of Fault Currents for Application of Power Circuit Breakers Rated on a Total-current Basis", ANSI Standard C37.5.

/4/ "Fast-Decoupled Load-Flow", B. Stott, O. Alsac, IEEE Transactions PAS, Vol. 93, pp. 859-869, 1974.

/5/ "A General-Purpose Version of the Fast-Decoupled Load-Flow", Robert A. M. van Amerongen, IEEE Transactions on PAS, Vol. 4, No. 2, May 1989.

/6/ "Optimal Power Flow Solutions", H. W. Dommel, W. F. Tinney, IEEE Transactions on PAS, Vol. 87, pp. 1866-1876, Oct. 1968.

/7/ "Optimal Load-Flow with Steady-State Security", IEEE Transactions on PAS, Vol. 93, pp. 745-751, May/June 1974.

/8/ "PAP; Power Systems Analysis Package", Project Report prepared by Electrical and Electronics Engineering Department, Middle East technical University, and submitted to Eregli Iron and Steel Works Inc., August 1989.

/9/ "Elektrik Güç Sistemleri Analiz Paketi (PAP)", O. Sevaioglu, First National Electrical Engineering Congress, pp. 258-266, Çukurova University, Adana, Turkey, October 25-27, 1985.

/10/ "Fast-Decoupled Load-Flow: Hypothesis, Derivations, and Testing", A. Monticelli, A. Garcia, O. R. Saavedra, Paper to be presented in the IEEE PES Winter Meeting, New York, New York, 1989.

/11/ "IDSPS, Elektrik Güç Sistemleri için Analiz Yazılımları Hazırlanması", Ç. Yalçındag, O. Alsaç, B. Dervişoğlu, O. Sevaioğlu, Project Report No. 75-04-01-01, Submitted to the Turkish Electricity Authority, prepared by the Electrical and Electronics Engineering Department, Middle East Technical University, 1976.

/12/ "Solution of the Multi-Machine Power System Stability Problem", C. Arnold, Ph. D. Thesis, University of Mancester, UK, 1976.

/13/ "IEEE Standard Application Guide for AC High Voltage Circuit Breakers Rated on Symmetrical Current Basis", IEEE Std. 320-1972, ANSI C 37.010-1972.

MODELLING AND SIMULATION OF CARRIER-MEDIATED ACTIVE TRANSPORT THROUGH LIQUID MEMBRANES

RADU MUTIHAC
Faculty of Physics, PO Box MG-11
University of Bucharest 76900
ROMANIA

and

LUCIA MUTIHAC
Romanian Academy
Institute of Physical Chemistry
Splaiul Independenţei 202
ROMANIA

Abstract- A mathematical model of ion pair active transport through an unsupported liquide membrane mediated by macrocyclic ligands as crown ethers and assisted by pH gradient has been developed. The overall mechanism is complex, consisting of both diffusion and complexation/decomplexation reaction steps at two possibly different interfaces. Modelling and simulation of these extreme processes (i.e., a diffusional regime and a kinetic regime), versus experimental data, make possible to determine qualitatively and quantitatively the rate limiting process, despite the poorly defined hydrodinamics of the cells involved in experiments. The main purpose of modelling and computer simulation was to optimize the rate of transport in amino acids separation techniques.

I. INTRODUCTION

The paper is focused on modelling ion pair transport through liquid membranes mediated by crown ethers and assisted by pH gradient. The possibility of separating amino acids from a mixture making use of macrocyclic ligands of crown ether type (18-crown-6) in a technique of simple extraction or two step extraction was previously reported (1). The macrocyclic ligands which are able to specifically recognize through controlled interactions according to size, shape and structure cations and then, by their means, various anions, has allowed the study of amino acids which amphion type molecule may appear either as cation in acid medium or as anion in basic medium. The possibility of coupling two more chemical species, according to a natural model by means of certain physico-chemical affinities without formation of covalent bonds is reffered to as "recognition" (2).

II. THE ION PAIR CARRIER MEDIATED TRANSPORT

This system has been most widely studied (3-5) especially in relation with the transport of bound monovalent cations (C^+) with an external anion(A^-) by neutral ligands(L) like crown ethers and cryptands. Generally, some of the phases are stirred to provide convective transport. The process is complex and the identification of rate-limiting process is complicated by the probable simultaneity of some steps and by other potential pathways involving complexation/decomplexation reactions in the bulk phases. Studies of the kinetics of substrate or ion pair selective transport may reveal the importance of different steps.

In order to increase the efficiency of the process an original design of the device was used(6). In the source phase there is the protonated amino acid(C^+) to be transported and the picrate anion(A^-) in an aqueous solution at pH=2 ensured by use of HCl. At the source phase/membrane interface, 18-crown-6 macrocyclic carrier complexes the amino acid through the protonated aminic group. The formed complex cation is extracted into 1,2 dichloroethane membrane by ion pairing with the picrate anion which diffuses through the membrane to the membrane/receiving phase interface. The receiving phase at pH=13, transforms the protonated aminic group - NH_3^+ into $-NH_2$; the amino acid is released from the complex and passes into receiving phase as a lithium salt. The empty carrier L diffuses back to the source phase/membrane interface where the whole cycle starts again and again until the equilibrium between the phases of the system is achieved. The basic pH of the receiving phase is ensured by using LiOH because the Li^+ ion, being much smaller than the crown ether ring size, is not complexed by the macrocyclic carrier and so it does not determine any additional reactions which could influence the transport phenomenon. Consequently, the acid medium of the source phase allows the formation of an ion pair complex(LC^+A^-) which diffuses through the membrane. The basic medium of the receiving phase facilitates the dissociation of the complex and thus the amino acid passes from the organic phase into the receiving phase. A coupled transport of amino acid and H^+ ions from source phase to receiving phase is accomplished.

II. TRANSPORT MODELLING

The overall process can be limited by the diffusion of the ion pair complex through the organic phase(LM) and /or by the rates of reactions occuring at the interfaces of the system. As long as the rates of complex formation and dissociation are fast compared to diffusion through the membrane, the amino acid flux is found to be:

$$J = PK[L^o]\frac{[C^+]_w^{in}[A^-]_w^{in} - [C^-]_w^{out}[A^-]_w^{out}}{(1+K[C^+]_w^{in}[A^-]_w^{in})(1+K[C^-]_w^{out}[A^-]_w^{out})} \quad (1)$$

where P is the permeation coefficient of the ternary complex and the free crown ether, K is the overall extraction equilibrium constant, and [ ] s are concentrations at the source phase/membrane (in) and the membrane/receiving phase(out) interfaces, respectively in aqueous(w) solutions. It is found that

$$J/J_{max} = \frac{a_1 K}{1+b_1 K+b_2 K^2} \quad (2)$$

and plotted versus lg K, where a_1, b_1 and b_2 are functions of the reaction evolution. Accordingly the reaction is most efficient (fast and selective) if lgK 4...6.

III. CONCLUSIONS

A diffusion-limited kinetic treatment and complexation/decomplexation limited processes have been modelled and computer simulated. Most experiments have suggested a diffusion-limited process.

CONVEX COMBINATIONS PROBLEM

AND

ITS APPLICATION FOR PROBLEM OF DESIGN OF LAMINATED COMPOSITE MATERIALS

A.G.Kolpakov
Orjonikidze st.,27,136,
630099 Novosibirsk,USSR

and

I.G.Kolpakova
NGPI and REKO,
Novosibirsk,USSR

Abstracts.
The paper is dealing with the convex combinations problem (CCP). Mathematical analysis of the CCP is presented. Results of the investigation are applied to problem of design of laminated composite materials.

I. CONVEX COMBINATIONS PROBLEM.

The CCP is formulated as follows:

find the set $L(\bar{Y})$ of all coefficients $(t_i, i=1,\dots,n)$ of convex combinations the given points $(\bar{Y}_i, i=1,\dots,m)$ which give the point $\bar{Y}$

or (which is the same)

find set of solutions of the problem

$$\sum_{i=1}^{m} \bar{Y}_i t_i = \bar{Y} , \quad t_i \geq 0 , \quad \sum_{i=1}^{m} t_i = 1 . \qquad (1)$$

The left hand side of the first equation in the (1) describes convex polyhedron $P=\mathrm{conv}(\bar{Y}_i, i=1,\dots,n)$. Let us consider set $(P_n, n=1,\dots,N)$ of simplexes of the polyhedron P containing the point $\bar{Y}$ (N is the total number of such symplexes) and denote by $\bar{S}_n(\bar{Y})$ solution of the CCP corresponding to the simplex P_n and point $\bar{Y}$.

Theorem 1. $L(\bar{Y})=\mathrm{conv}(\bar{S}_n, n=1,\dots,N)$ (conv denotes the "convex hull").

Let us define distance between two polyhedrons A,B by the formula

$$D(A,B) = \max_{\bar{y}\in A\setminus B} \max_{\bar{x}\in B} \left| \bar{x} - \bar{y} \right| + \max_{\bar{y}\in A\setminus B} \max_{\bar{x}\in B} \left| \bar{x} - \bar{y} \right| \qquad (2)$$

Theorem 2. Let $\bar{Y}_a \to \bar{Y}$ when $a \to 0$ and $\bar{Y}_a, \bar{Y}$ belong to P for every a. Then $D(L(\bar{Y}), L(\bar{Y})) \to 0$ as $a \to 0$.

II. NUMERICAL ALGORITHM.

As seen, solution of the CCP is reduced to the finding of the finite number of solutions $(\bar{S}_n, n=1,\dots,N)$. For 1-D problem these solutions can be computated explicitly /1/. Let us consider the first equation in the CCP (1) as 1-D CCP:

$$\sum_{i=1}^{m} Y_{i1} t_i = Y_1 , \quad t_i \geq 0 , \quad \sum_{i=1}^{m} t_i = 1 . \qquad (3)$$

In accordance with the theorem 1, its solution is

$$t_i = \sum_{n_1=1}^{N_1} S_{n_1 i} u_{n_1} , \qquad (4)$$

where

$$t_{n_1} \geq 0 , \quad \sum_{n_1=1}^{N_1} t_{n_1} = 1 . \qquad (5)$$

Substituting (4) into the equations from the 2d to the m-th in (1) one obtains a new 1-D CCP (with respect to $(u_{n_1}, n_1=1,\dots,N_1)$. After m steps one obtains solution of the CCP (if it is solvable) in the form

$$t_i = \sum_{n_m=1}^{N_m} H_{n_m i} u_{n_m} ,$$

where

$$u_{n_m} \geq 0 , \quad \sum_{n_m=1}^{N_m} u_{n_m} = 1 ,$$

and

$$H_{n_m 1} = \sum_{n_1=1}^{N_1} \dots \sum_{n_{m-1}=1}^{N_1} s_{n_m} s_{n_{m-1}} \dots s_{n_1 i} .$$

If the CCP (1) has no solution, then a 1-D CCP in the algorithm above is not solvable.

III. APPLICATION OF THE CCP FOR DESIGN OF COMPOSITE MATERIALS OF LAMINATED STRUCTURE WITH SPECIFIED PROPERTIES.

Relationship between homogenized characteristics /2/ $\bar{A}$ of composites of laminated structure and characteristics of their components has the following form /3-6/:

$$\bar{G}(\bar{A}) = \sum_{i=1}^{m} \bar{F}(\bar{U}_i) t_i \ , \qquad (6)$$

where:
$\bar{U}_i$ denotes mechanical characteristics of the i-th material /4/ (or direction of fibers of the i-th layer /5/);
t_i is the volume ratio of the i-th material (or fibers in the i-th layer);
m is the total number of materials or layers of fibers which may be used;
$\bar{G}$ and $\bar{F}$ are known functions.
In accordance the with the definition of $(t_i, i=1,\ldots,m)$ one has

$$t_i \geq 0 \ , \quad \sum_{i=1}^{m} t_i = 1 \ . \qquad (7)$$

Problem of design of composite with specified properties is formulated in the following form /4/:

find a structure of composite (it is described by $(t_i, i=1,\ldots,m)$) which gives to composite a set of required homogenized characteristics $\bar{A}$

or (which is the same)

solve problem (6),(7) with respect to $(t_i, i=1,\ldots,m)$ when $\bar{A}$ is given.

As seen, the problem of design becomes a CCP.
The numerical algorithm above was realized as FORTRAN program. It demonstrated its efficiency for number of materials up to 50 and number of homogenized characteristics up to 10.

The problem of design of composite constructions of laminated structure with specified characteristics is as follows:

find a structure of a composite which gives to the composite a required distribution of homogenized characteristics $\bar{V}(\bar{x})$ in the domain Q /7/.

It comes to a CPP with right hand side $\bar{V}(\bar{x})$ (where $\bar{x} \in Q$ is a parameter), which can be approximated by finite number of CPPs with right hand sides $\bar{V}_l = \bar{V}(\bar{x}_l)$, $l=1,\ldots,L$ on the base of theorem 2.

REFERENCES

1. Kolpakov A.G. Usrednennie harakteristici sloistih compositov (chislennii algoritm) // Chislennii metodi reshenia zadach uprugosti i plastichnosti. Materiali IX Vses.konfer. Novosibirsk: ITPM SO AN SSSR, 1986. P.163-165 (in Russian).
2. Bensoussan A., Lions J.L., Papanicolaou G. Asymptotic analysis for periodic structures. Amsterdam: North-Holl. Publ. Comp. 1978.
3. Sanchez-Palencia E. Non-homogeneous media and vibration theory. Lect.Notes in Phys.127, Springer-Verlag,1980.
4. Kolpakov A.G. Design of Laminated Composites// Proceedings of the 5-th Intl Symposium on Numerical Methods in Engineering. Vol.1. Computational Mechanics Publ., 1989.
5. Annin B.D., Kalamkarov A.L., Kolpakov A.G. Analysis of Local Stressses in Highmodulus Filamentary Composites// Localized Damage Computer-Aided Assessment and Control. Vol.2. Computational Mechanics Publ.1990.
6. Kolpakov A.G., Kolpakova I.G. Teploprovodnost cherez granizu s nelocalnimi characteristicami// Differencialnii uravnenia, 1985,t.21,N9 (in Russian).
7. Alehin V.V., Annin B.D., Kolpakov A.G. Syntez sloistih materialov i constructii. Novosibirsk: IG SO AN SSSR, 1988 (in Russian).

PROFIT. A COMPUTER PROGRAM FOR SELECTING AND FITTING PROBABILITY DISTRIBUTIONS TO DATA

J.A. Taylor[1], J. Bai[1,2], A.J. Jakeman[1], and M. McAleer[3]

1. Centre for Resource and Environmental Studies, Australian National University, GPO Box 4, Canberra, ACT, 2601, Australia.

2. Statistics Research Section, School of Mathematical Sciences, Australian National University, GPO Box 4, Canberra, ACT, 2601, Australia.

3 Department of Economics, University of Western Australia, Nedlands, Western Australia, 6009.

Abstract - A computer program for fitting and discriminating between probability distributions has been developed to run on PCs. The program is for those interested in fitting distributions to data, for example, environmental, life testing, meteorological and traffic count data. The program uses advanced state-of-the-art discrimination and maximum likelihood parameter estimation techniques and these have been applied to a wide range of practical problems.

Introduction

Based on extensive research into the application of probability distributions to environmental data, a computer program called PROFIT has been developed. Environmental data can usefully be summarised using probability distributions as an aid to analysis and decision making. In many cases the most useful environmental indicators are the mean, variability, extreme events and the probability of exceeding specified values. Such indicators form the basis for most environmental standards. All these indicators of environmental quality can be derived from the correctly identified probability distributions and their parameter values. PROFIT is a flexible program for applying probability distribution models to data. It combines powerful identification tools with maximum likelihood parameter estimation in an easy to use package.

Overview

The program is specially designed to be used by those in industry, government and research who have a special interest in fitting probability distributions to data. The program can be usefully applied to those wishing to study data sets arising from reliability and quality control tests, air and water quality data sets, meteorological data and road traffic data. However, application of the program is really only limited by the applicability of the distribution models included in the program to the data in question.

In order to simplify the program structure we have selected the most widely applicable probability distributions such as the gamma, Weibull, lognormal, normal and exponential models. For the gamma, Weibull and lognormal distributions, two- and three-parameter distributions and the ability to select between these distributions have been included in the program.

While many computer programs used for statistical analysis of data are available, no other program offers such ease of use, flexibility and advanced state-of-the-art techniques with extensive proven application to a wide range of real data sets.

Program Features

The program incorporates many useful design features based upon the authors own experiences with the analysis of environmental data sets. These design features make the program powerful and easy to use. They include:

(i) Flexible command structure allowing both interactive and batch processing of data. The analysis of individual data sets can be rapidly achieved using the program in an interactive mode. This feature is particularly useful when conducting an exploratory analysis. However, where a standard set of outputs are required or multiple data sets must be analysed then batch files can be input to the processor.

(ii) Easy to use interfaces based on menus makes the program easy to learn and easy to use.

(iii) Straightforward formats for data input speed analysis. A range of simple unformatted data files can be accepted by default. Alternatively, to assist with the development of data bases, a simple hierarchal data format has also been developed. Unformatted data can be written in the hierarchal format so that no preprocessing is required to take full advantage of this feature.

(iv) An on-line help facility is available, with examples, which provides user selected levels of help for the new and experienced user. Program generated error messages provide detailed descriptions of problems and the most likely remedies.

(v) The processors will produce a set of default or standard outputs User definable outputs can be easily added.

(vi) A flexible plotting utility is available to aid in the analysis of the input data and the presentation of results. These include plots of autocorrelation, time series, error analysis and time series plots to aid in the assessment of stationarity.

(vii) A selection of widely applicable distributions have been included. These are the normal, the exponential distribution, two- and three-parameter Weibull distributions, and two- and three-parameter lognormal distributions.

(viii) A key processor designed to perform the maximum likelihood estimation of parameter estimates for the probability distribution models has been designed to be applicable to the full range of potential parameter values for each distribution. The parameter estimation processor has been tested on a wide range of actual environmental data sets.

(ix) A comprehensive range of well tested state-of-the-art identification and selection criteria have been incorporated into the program. A special rule based expert system aids in the interpretation of the results.

(x) The program also includes simulation tools to enable sampling from the probability distributions listed in (vii) above. The simulation tool allows quantitative investigation of the ability of the identification and parameter estimation procedures to predict desired properties or derived quantities of probability distributions. This facility allows, for example, the user to

investigate the ability of the maximum likelihood approach to estimate a percentile of a particular distribution.

(xi) Another important simulation tool is the application of the bootstrap method to data sets in order to determine the variability in a desired user specified quantity such as a percentile or parameter value.

Distribution Functions, Statistical and Performance Criteria and Asymptotic Tests

I. Distribution functions

For the exponential and normal distributions the reader is referred to the extensive literature available on these distributions [1].

Probability density functions for the three-parameter gamma, Weibull and lognormal distributions for a random sample are given by:

Gamma:
$$f(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\gamma}{\beta}\right)\right] \quad (1)$$

Weibull:
$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right] \quad (2)$$

Lognormal:
$$f(x) = \frac{1}{\alpha\sqrt{2\pi}}(x-\gamma)^{-1}\exp\left\{-\frac{[\log(x-\gamma)-\beta]^2}{2\alpha^2}\right\} \quad (3)$$

In equations (1), (2) and (3), α represents the shape parameter, β the scale parameter, γ the location parameter, and Γ is the gamma function. The two-parameter versions of the density functions of the gamma, Weibull and lognormal distributions are the same as in (1), (2) and (3), with $\gamma = 0$ in each case. In the above equations, $\beta > 0, \alpha > 0$ and γ is less than the minimum observed sample value.

The maximised value of the likelihood function is an essential statistic employed in many criteria used to discriminate among alternative models. For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the log-likelihood functions for the three-parameter gamma, Weibull and lognormal distributions are given as follows:

Gamma:
$$\log L = -n\alpha\log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log(x_i-\gamma) - \sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right) \quad (4)$$

Weibull:
$$\log L = n\log\alpha - n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log(x_i-\gamma) - \sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \quad (5)$$

Lognormal:
$$\log L = -\frac{n}{2}\log(2\pi\alpha^2) - \sum_{i=1}^{n}\log(x_i-\gamma) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}[\log(x_i-\gamma)-\beta]^2 \quad (6)$$

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure for the three-parameter gamma and Weibull distributions will frequently fail to converge when the (unknown) shape parameter is less than or equal to unity, a computationally efficient approach that circumvents this problem is used (for further details, see [2]).

II. Statistical Criteria

Let $x_1, x_2, \ldots, x_n$ represent a random sample of n observations. Interest here lies in discriminating among nested and non-nested two- and three-parameter distributions in which the null hypothesis of interest is $H_0: \gamma = 0$ against the alternative $H_1: \gamma \neq 0$. Denoting the maximised values of the two- and three-parameter variants of a particular log-likelihood function as $\log L_0$ and $\log L_1$, respectively, the AIC and SIC may be expressed, respectively, as:

Choose the $\left\{\begin{matrix}2\\3\end{matrix}\right\}$ parameter distribution if

$$AIC: \log L_0 - 2\left\{\begin{matrix}>\\<\end{matrix}\right\}\log L_1 - 3 \quad (7)$$

$$SIC: \log L_0 - \log n\left\{\begin{matrix}>\\<\end{matrix}\right\}\log L_1 - 3\log n/2 \quad (8)$$

A new GIC procedure is proposed for the discrimination of distributional structures among a set of alternatives by Bai et al. [3]. The GIC is based on the equivalence between some well-known information criteria and hypothesis tests, and attempts to determine the false distributions based on sample information. Large differences between the maximised values of log-likelihood functions will lead to rejection of the distribution with the lower value. Discriminated distributions are separated into two categories, the superior and badly fitting categories. The distributions in the superior category perform within an acceptable tolerance level and there are no significant differences among their performances. The GIC procedure may provide several alternatives rather than one particular distribution for a particular set of data. In the event of there being several sets of data, the distribution with the highest probability of acceptance in the superior category will be chosen.

Let $x_1, x_2, \ldots, x_n$ be n independently and identically distributed random observations. Denote $\log L_j$ as the maximised log-likelihood value of distribution j (j=1, 2, ..., m), with the ordering given as

$$\log L_1 > \log L_2 > \cdots > \log L_m \quad (9)$$

Then distribution j will be rejected in favour of distribution 1 if

$$GIC: -2\log L_1 + T_0 < -2\log L_j \quad (10)$$

where $T_0 > 0$. The value T_0 is the tolerance level required in order to reject distributions as being significantly different from each other, and is equivalent to the rejection region discussed in the previous section. For the nested case, T_0 can be expressed as

$$T_0 = c \quad (11)$$

where c is the critical value of the χ^2 distribution with one degree of freedom. In the non-nested case, T_0 is given by

$$T_0 = 2z^* \quad (12)$$

where $z^* > 0$ and the asymptotic upper bound on the significance level is given by the cumulative standard normal distribution function evaluated at $\phi\{-\sqrt{2}z^*\}$.

The motivation behind the GIC procedure is straightforward. First, for a given sample, select the distribution with the highest maximised log-likelihood value among $\log L_j$ (j = 1, 2, ..., m). This distribution then belongs to the category of superior distributions and is also used as a standard for further inference. Second, reject the false alternatives among the remaining distributions in terms of the given tolerance. The distributions which perform within an acceptable tolerance level of the best fitting distribution are retained in the superior category, and the distribution with the highest probability of acceptance over different sets of data will be chosen from the superior category.

III. Performance Criteria

In this study, loss functions recommended for assessing air quality models have been chosen (see [4]) to establish the effect of discrimination criteria on the intended use of the model. These functions are the relative bias (BIAS) and the relative root mean square error (RRMSE) which are evaluated at the upper percentiles of the distributions. For an estimate $\hat{q}_i$ of a quantity of interest q, these loss function are defined in terms of deviations from the observed value q in each data set. The definitions used for the loss functions are:

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{q}_i - q}{q}\right) \tag{13}$$

$$RRMSE(q) = \left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{q}_i - q}{q}\right)^2\right\}^{0.5} \tag{14}$$

where N is the number of data sets. For present purposes, the quantity q denotes the upper percentiles of the sampling distributions.

Based on the loss functions above, two performance criteria defined in terms of the relative root mean square error (RRMSE) are also used. The first criterion is based on the mean of RRMSE in the upper (U) percentiles, and the second is based on the mean of RRMSE in the entire or full (F) percentiles of the distribution. For an estimate $\hat{q}_{ij}$ of a quantity of interest q, these performance criteria are defined in terms of deviations from q in the percentiles of interest, where i denotes the specific percentiles estimated and j corresponds to the specific data set fitted. The definitions of the upper percentile error and full percentile error are as follows:

$$UPE = \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{1+(1-p)n_j}\sum_{i=pn_j}^{n_j}\left(\frac{\hat{q}_{ij} - q}{q}\right)^2\right\}^{0.5} \tag{15}$$

$$FPE = \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}\left(\frac{\hat{q}_{ij} - q}{q}\right)^2\right\}^{0.5} \tag{16}$$

where p is the location of the p-quantile which is chosen at the 98 per cent level for UPE and 100 per cent for FPE, N is the number of data sets, and n_j is the sample size for data set j. For present purposes, q denotes the percentile quantities related to the upper and full percentile errors.

IV. Asymptotic Tests

The performance of two well known procedures for testing goodness of fit are also considered, namely the chi-square (CHI) test and Kolmogorov-Smirnov (KS) test. Classifying the n observations into k categories, the chi-square statistic is of the form (see [5]):

$$CHI = \sum_{i=1}^{k}\frac{(f_i - np_i)^2}{np_i} \tag{17}$$

which has an asymptotic χ^2 distribution with $(k-l-1)$ degrees of freedom when H_0 holds. The p_i are hypothetical probabilities, the f_i are empirical frequencies and l is the number of parameters estimated for each distribution (for further details, see [1]). For fitting the real data in Sections 5 and 6 below, $k = 10$ and $l = 2$ or $l = 3$. The KS test, which is defined in terms of the maximum absolute difference between the sample distribution function $S_n(x)$ and the hypothetical distribution function $F_0(x)$ (see e.g. [6], p. 204), is given by

$$D_n = \sup_x |S_n(x) - F_0(x)|. \tag{18}$$

Large observed values of the D_n statistic lead to rejection of the hypothesis $F_0(x)$.

An Example Application: Chemical Determinants in Stream Water

Jakeman et al [7] developed a hybrid modelling approach to predict extreme concentrations of chemical determinants in stream water. Jakeman et al [7] used the PROFIT computer program to identify and estimate the parameters of the probability density function (pdf) component of the hybrid model. We provide a brief description of the study reported by Jakeman et al [7] illustrating the application of the PROFIT computer program.

The water quality data used here comprise weekly observations on stream water chemical determinands from the Birkenes catchment in Norway. There are fourteen years of reasonably complete records available for five chemical species: sulphate, nitrate, sodium, calcium and hydrogen ions. The parameter distributions applied to characterise these data are the 2- and 3-parameter gamma (G2 and G3), lognormal (L2 and L3) and Weibull (W2 and W3). As reviewed in Jakeman and Taylor [8] these distributions are generally capable of representing skewed environmental data.

Before embarking upon the analysis, it is important to appreciate that an underlying assumption is theoretically required when pdfs are fitted to observations using techniques such as maximum likelihood. This assumption is that the observations are independent and identically distributed (iid) over the sampling period, i.e. stationarity and low autocorrelation for observations. To obtain an indication of the satisfaction of this assumption, three simple analyses were undertaken. First, the variability of the mean between the summer and winter seasons was examined for each annual period. Large differences for most years would imply a non-stationary seasonality for the variable and hence the need to build a causal model with seasonal rather than annual predictive capabilities. Second, regressions against flow were constructed for the different variables for each year. This would indicate if short term concentrations were flow related and hence if they might be better modelled taking this variable into account. The regressions were performed for concentration on flow and for the logarithm of these two variables, but invariably the latter produced higher levels of explanation. Third, autocorrelation functions were computed for each variable and for each year.

A summary of the outcome of these exercises, using the median result, is given in Table 1. The autocorrelations are high. However, Jakeman et al [9] report that, for the L2 distribution, first-lag autocorrelations of 0.5 lead to increases in fitting errors for the maximum value of 25 per cent. This increase is relative to random samples of the same size from the same lognormal distribution. It could be regarded as a worst case due to the larger positive skewness of the lognormal distribution over other distributions used in this paper. The table shows that both calcium and hydrogen have substantial correlation with flow. There is little doubt that short term hydrogen concentrations should be explained by flow and possibly other variables. To a lesser extent, this is the situation for calcium as well. In the remainder of this paper, we will therefore restrict attention to the sulphate, nitrogen and sodium species. The initial analyses reported in Table 1 indicate that annual sets of short term concentrations of these species are suited to modelling by pdfs. For nitrate there does exist almost 30 per cent difference between the summer and winter means for half of the annual data sets. As we shall see, the accuracy of our statistical modelling of extremes of nitrate is lower than that for sodium or sulphate.

I. Estimation of distributional properties

Tables 2, 3 and 4 report the results of fitting by maximum likelihood the six distributional forms to the 14 annual sets of weekly observations of sulphate, nitrate and sodium concentrations, respectively. The tables give statistics associated with the rrmse in fitting the observed mean and the observed maximum. The three statistics shown are the average of the rrmse over the 14 data sets, the standard deviation and the maximum rrmse computed.

TABLE 1 Summary of Investigation of IID Assumptions on Stream Acidity Variables at Birkenes

Species	Median value over 14 years		
	Relative difference in summer Vs winter mean	R^2 correlation with flow*	First lag autocorrelation
sulphate	0.08	0 19	0 66
nitrate	0 29	0 09	0 55
sodium	0.12	0 06	0 65
calcium	0.17	0 45	0 64
hydrogen	0.23	0.57	0 51

*correlation is result of regression of log species concentration on log flow, which is generally higher than that of concentration on flow.

TABLE 2 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Sulphate Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0 017	0 017	0 070	0 079	0 050	0.183	10	1	1	8	10
G2	0 006	0.002	0 008	0.064	0 046	0.178	9	4	4	9	14
L3	0 015	0 010	0 041	0 074	0 039	0.140	7	1	1	7	8
L2*	0.009	0 003	0.014	0 065	0 048	0.176	12	3	5	9	14
W3	0 024	0 033	0.137	0 063	0 042	0.139	12	3	1	8	13
W2	0 361	0 475	0.999	X	X	X	2	2	2	5	7

X = large error
* = preferred distribution
AVG = average of rmse over 14 data sets
SDV = standard deviation
MAX = maximum of rmse values

GIC = Generalised Information Criterion
AIC = Akaike Information Criterion
SIC = Schwarz Information Criterion
CHI = Chi square test
K-S = Kolmogorov-Smirnov test

TABLE 3 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Nitrate Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0.252	0.206	0.641	0.364	0 378	1.363	10	6	6	5	6
G2	0.130	0.066	0.263	0.116	0 089	0.312	8	4	4	12	14
L3	0.165	0.139	0.548	77.497	278.706	1082.385	0	0	0	1	1
L2	0.192	0 096	0.380	0.493	0.494	2.021	5	0	0	8	13
W3	0.205	0.187	0.594	0 696	1.094	3 041	9	0	0	8	9
W2*	0 099	0.074	0.265	0.102	0 074	0.288	8	4	4	10	14

TABLE 4 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Sodium Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0 013	0 011	0 038	0.056	0 063	0.245	8	1	1	7	8
G2*	0 006	0 004	0 013	0 047	0.067	0 284	12	6	7	11	14
L3	0 012	0.010	0 042	0 049	0 053	0.226	7	1	1	7	7
L2	0 008	0 006	0 020	0 050	0 068	0.284	12	3	5	12	14
W3	0 030	0 042	0.133	0 057	0 059	0 230	11	3	0	10	12
W2	0 576	0 489	1 000	X	X	X	3	0	0	4	5

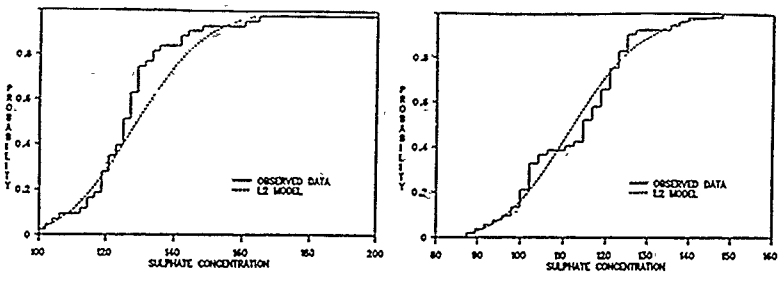


Fig. 1. Fit of the L2 distribution to weekly sulphate concentrations in 1986 and 1987.

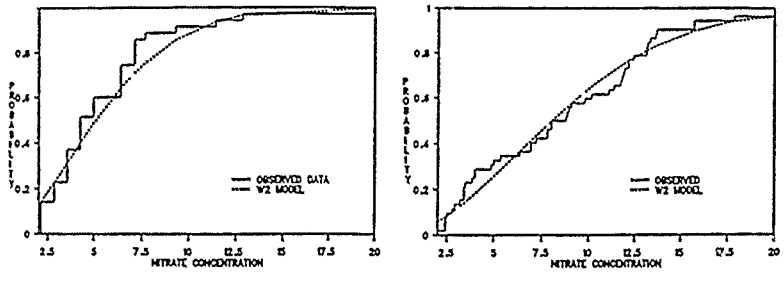


Fig. 2. Fit of the W2 distribution to weekly nitrate concentrations in 1980 and 1987.

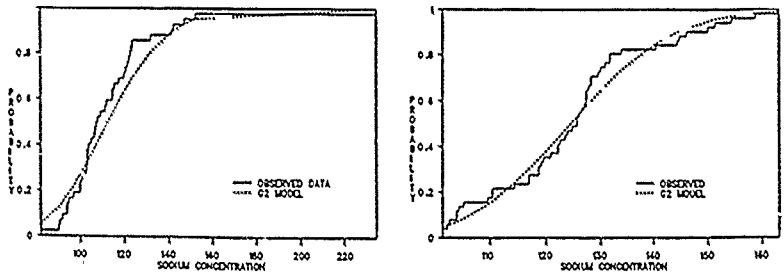


Fig. 3. Fit of the G2 distribution to weekly sodium concentrations in 1988 and 1983.

In the case of sulphate, the L2 distribution fits the observed mean concentration with an accuracy of 0 9 per cent on average and, in the worst case, the error of fit is 1.4 per cent on one of the 14 data sets. The L2 distribution fits the observed maximum of each annual data set with an average accuracy of 6.5 per cent, the worst fit over all years being 17.6 per cent in error. Only the W2 distribution yields unacceptable errors in fitting these quantities.

For nitrate, all 3-parameter distributions and L2 yield unacceptable fits to the observed maximum. W2 is the most superior distribution but only marginally to G2. For both these distributions, the average error in fitting the observed mean and maximum is close to 10 per cent and the worst error on any one data set is around 30 per cent.

For weekly sodium concentrations, only the W2 distribution yields poor fitting errors. The 3-parameter distributions provide no substantial improvements over the more parsimonious alternatives. The G2 and L2 distributions fit the observed mean very closely and better than the others, while the observed maximum is fitted to an accuracy of less than 5 per cent on average and always less than 30 per cent.

II. Discrimination among distributions

Tables 2, 3 and 4 also show the results of applying the GIC, AIC and SIC statistics, and the traditional chi-square and K-S tests. All tests apply a 98 per cent confidence level. Consider the sulphate results first. The GIC selects L2 and W3 as not significantly inferior to alternatives in 12 of the 14 cases. The other two information criteria (AIC and SIC) are not as definitive. The chi-square test and especially the K-S test are more definitive preferring G2 and L2, followed by W3 and G3. L2 is certainly the most preferred or equally preferred by all criteria except AIC. The error results in Table 2 only eliminate W2. If a 2-parameter distribution is sought, then either L2 or G2 would suffice, our preference being L2 because of its likelihood (by virtue of GIC results) being optimal or near optimal most often.

For nitrate concentrations, Table 6 shows that all the information criteria prefer the G3 distribution. The chi-square test prefers G2 while the K-S test finds all 2-parameter distributions acceptable. On the basis of the errors of fitting the observed means and maxima, as reported in Table 3, only G2 and W2 can be considered sufficiently accurate. The K-S and chi-square tests rate these most highly also while the information criteria rank them reasonably highly. The W2 distribution seems marginally preferable on fitting error grounds

The results of applying the three information criteria and two asymptotic tests for sodium concentrations are quite consistent. The G2 distribution is most preferred or at least equally preferred to L2 by four of the five discrimination methods, the chi-square test accepting G2 for 11 of the 14 times and L2 on 12 occasions. The G2 distribution also fitted the observed means most closely and on average fitted the observed maxima with least error. However, the L2 distribution appears almost as acceptable as G2.

Fig. 1 illustrates the range of the fitting performance attained by the L2 distribution when estimation is by maximum likelihood on annual sets of weekly sulphate concentrations. The left hand plot shows the worst fit to the observed maximum concentration obtained over the 14 data sets (in 1986) while the right hand plot shows the best fit to this observation (in 1987). Figs. 2 and 3 repeat this illustration of the range of error in fitting the observed maximum for nitrate and sodium. In each case, our preferred distribution is used, the W2 and G2 distributions, respectively.

Conclusions

A computer program for selecting and fitting probability distribution functions to data has been developed which offers the advantages of being easy to use, flexible and incorporating state-of-the-art algorithms and numerical techniques The program has been successfully applied to a large range of data sets, as illustrated in the example application. The program can be applied to any data set to which the normal, exponential and 2- and 3-parameter Weibull, gamma and lognormal probability distributions are applicable.

References

[1] M.G. Kendall, and A. Stuart, *Advanced Theory of Statistics*, Vol.2(4), 1979.

[2] J. Bai, A.J. Jakeman and M. McAleer, *A new approach to maximum likelihood estimation of the three-parameter gamma and Weibull distributions*, Working Paper in Economics and Econometrics, No. 191, Australian National University, pp. 26, 1989.

[3] J Bai, A.J Jakeman and M. McAleer, *Discrimination procedures for fitting nested and non-nested distributions to environmental quality data*, Working Paper in Economics and Econometrics, No. 200, Australian National University, pp. 58, 1990

[4] D G Fox, 'Judging air quality model performance', *Bulletin of the American Meteorological Society*, Vol. 62, pp 599-609, 1981

[5] E S Pearson, 'Note on an approximation to the distribution of non-central χ^2', *Biometrika*, 46, 364, 1959.

[6] K.V. Bury, *'Statistical Models in Applied Science'*, John Wiley, New York, 1975.

[7] A.J. Jakeman, P.G Whitehead, A Robson, J A Taylor and J Bai, 'Investigation of a new approach to predict water quality extremes with a case study of chemical determinants in stream water', presented at Water Quality, Baltimore, 1991.

[8] A J. Jakeman and J A. Taylor, 'Identification, estimation and simulation of frequency distributions of pollutant concentrations for air quality management'. In: *Library of Environmental Control Technology*, P N Cheremisinoff (ed), Gulf Publishing, Houston, 2, 135-158, 1989.

[9] A.J Jakeman, R.W Simpson, and J A. Taylor, 'Modelling distributions of air pollutant concentrations - III The hybrid deterministic-statistical distribution approach' *Atmos Env.*, 22, 163-174, 1988.

FEATURES AND APPLICATIONS OF IHACRES, A PC PROGRAM FOR IDENTIFICATION OF UNIT HYDROGRAPHS AND COMPONENT FLOWS FROM RAINFALL, EVAPOTRANSPIRATION AND STREAMFLOW DATA

A J Jakeman[1], I G Littlewood[2], H D Sykes[1]

1. Centre for Resource and Environmental Studies, Australian National University, GPO Box 4, Canberra, ACT, 2601, Australia.
2. Institute of Hydrology, Wallingford, OX10 8BB, United Kingdom.

Abstract - A methodology for identification of unit hydrographs corresponding to total river flow has been developed [1] and programmed for PC usage. The underlying mathematical model employed to convert rainfall into streamflow is presented. Major products and features of the associated PC package, IHACRES, are described using examples from various applications of the methodology to hydrological systems where either hourly or daily rainfall and streamflow data have been available.

Introduction

Unit hydrograph theory [2] plays a central role in many established numerical procedures for characterising the dynamics of the rainfall - streamflow process at catchment scale. In many procedures, an initial step is to subtract baseflow from the observed hydrograph in some intuitively reasonable but fairly arbitrary way to give a residual streamflow. Then, a unit hydrograph (impulse response function) is identified which links the system input of time variations in rainfall excess to the system output of time variations in residual streamflow. The methods presently used for adjusting rainfall to give rainfall excess, and any necessary parameters, are chosen or optimised to give a best fit between 'observed' and modelled residual streamflow. A disadvantage of this overall approach is that the derived unit hydrograph applies only to a rather poorly defined component of streamflow (i.e. observed streamflow minus a conjectural baseflow). Although a unit hydrograph derived in this way might characterise well the dynamic relationship between rainfall and streamflow for the individual 'well-behaved' event on which the model was calibrated, it may perform rather poorly in validation mode on other events. It is common practice with such methods, therefore, to derive an average unit hydrograph for a catchment from calibrations on several specially selected events (e.g. events which exhibit a dominant peak and where the rainfall has a 'well-behaved' temporal profile).

In [1], a methodology is presented for deriving unit hydrographs corresponding to (a) total streamflow and (b) the quick and slow components of streamflow separately. This methodology has been programmed for PCs and the package has been called IHACRES (Identification of unit Hydrographs And Component flows from Rainfall, Evaporation and Streamflow data). IHACRES employs time series of rainfall and streamflow data for model parameter estimation. It does not require selection of 'well-behaved' events. Neither is separation of baseflow from the hydrograph required before unit hydrograph identification. From the high quality of calibration and validation model-fits which have been obtained [1], [3-6] for different catchment types and data time intervals (e.g. hourly, daily), it is evident that unit hydrographs can be well identified by IHACRES. The identification of separate unit hydrographs for quick and slow flow provides the facility to separate stream hydrographs into quick and slow flow components.

The capability of IHACRES to derive reliable unit hydrographs for total streamflow, and for flow components (quick and slow) separately, makes it widely applicable in studies of hydrological systems. In connection with several types of investigation of environmental systems, IHACRES has already been applied to rainfall and streamflow data sets from the United Kingdom, Australia, New Zealand and the United States of America. Jakeman, Littlewood and Whitehead [4] use the package to assess the catchment-scale hydrological impact of partial-area land-use changes associated with the forestry cycle in two catchments in Central Scotland. Littlewood (unpublished data) and Jakeman, Littlewood and Whitehead [1] investigate the relationship between the hydrological characteristics and dynamics of surface water acidity for small upland catchments in Wales. Jakeman and Littlewood [5] investigate the change in hydrological characteristics of a 20 km catchment in the Australian Capital Territory over a period of mature eucalypt forest regeneration after a severe bush fire. Littlewood and Jakeman [6] assess the temporal invariance of the unit hydrograph over a 30 year period for a large (894 km²) catchment in Wales. Data from several small research catchments around the world are being analysed currently by IHACRES to examine the hypothesis that peak streamflows can comprise mostly 'old' water which is displaced by 'new' rainwater (see later in this paper). Other possible applications of IHACRES related to engineering hydrology and water quality hydrology are being considered.

This paper introduces the PC package, IHACRES, developed jointly by the Institute of Hydrology (IH) and the Centre for Resource and Environmental Studies (CRES); the name 'IHACRES' also reflects the collaboration to this end between IH and CRES. The paper begins with details of the model invoked by IHACRES and then illustrates various features of the package, employing selected examples from case studies undertaken by the authors.

The model

The model used in IHACRES is based on unit hydrograph theory (e.g. [2]) which describes the variation of streamflow $x(t)$ over time t as a linear convolution between rainfall excess $u(t)$ and the unit hydrograph $h(t)$, viz.

$$x(t) = \int_0^t h(t-s)\, u(s)\, ds \qquad (1)$$

Rainfall excess is the amount of rainfall which contributes to streamflow after 'losses' due to evapotranspiration have been deducted. When dealing with a continuous time representation, as in (1), the unit hydrograph is the streamflow response to a unit of rainfall excess applied instantaneously. When a discretisation of (1) is used, as in IHACRES, the corresponding input is unit rainfall excess over one sampling interval.

The assumptions of unit hydrograph theory are well known, but are conveniently summarised in [1]. The methodology in IHACRES also invokes an approximation of the unit hydrograph as a combination of exponential decays. Another interpretation of this is that rainfall excess is input to a system of linear reservoirs or storages from which there is an output which represents streamflow. The configuration of reservoirs may involve storages that are in parallel, in series or both. User-interaction with IHACRES can identify the most appropriate configuration in each case (see [1]); the most common configuration encountered by the authors is two storages in parallel. When derived in discrete-time, and in the case of two storages acting in parallel, the relationship between rainfall excess u_k and streamflow x_k at time step k can be written as a second-order transfer function of the form

$$\left.\begin{aligned} x_k &= x_k^q + x_k^s \\ x_k^q &= \beta_q\, u_k - \alpha_q\, x_{k-1}^q \\ x_k^s &= \beta_s\, u_k - \alpha_s\, x_{k-1}^s \end{aligned}\right\} \qquad (2)$$

(We designate the parameters and streamflow component variable of the storage with the quicker throughput with sub- or superscript q, whereas the slower storage properties are denoted with an s). The model is shown schematically in Fig. 1. Each storage can be defined alternatively by any two of the following three parameters which we call the characteristic catchment properties: the time constant (τ), relative throughput volume (V) and contribution to the peak of the unit hydrograph (I). These characteristic catchment properties are defined as:

$$\left.\begin{aligned} \tau &= -\Delta t/\ln(-\alpha) \\ gV &= \beta/(1+\alpha) \\ gI &= \beta \end{aligned}\right\} \qquad (3)$$

where g and p see [1] are unity if the volume of rainfall excess is equated to streamflow over the calibration period by taking into account the catchment area and the units of measurement. Δ is the sampling interval for the time series of rainfall and streamflow used in the parameter estimation.

interaction with IHACRES. When modelling with daily data over long periods during which there may be seasonal changes in evapotranspiration, the rainfall can be adjusted first according to the difference between the mean air temperature for the month and an overall maximum temperature determined by trial and error.

Thus

$$r_k^* = \left(1 - \frac{t_k}{t_m}\right) r_k \qquad (4)$$

where t_m is a reference temperature greater than the recorded maximum for the location in question and, for calculating r_k^* in any given month, t_k is the observed mean temperature for that month.

A catchment wetness index, s_k, is calculated according to

$$s_k = s_{k-1} + \tau_w^{-1}\left(r_k^* - s_{k-1}\right) \qquad (5)$$

where τ_w is a constant related to the time of wetting up of a catchment.

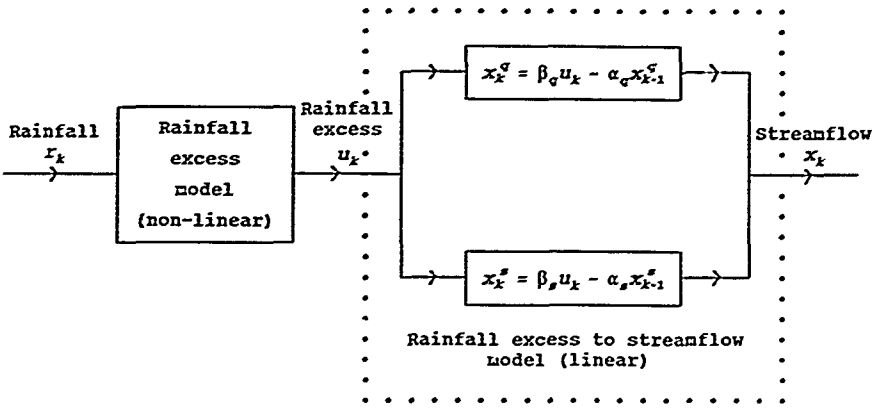


Fig. 1. Systems diagram of the most common model configuration used by IHACRES.

The package also allows use of a non-linear model of the relationship between rainfall and rainfall excess. Rainfall excess is highly dependent on the level of antecedent rainfall and may also depend on changes in evapotranspiration. The model used in IHACRES is a simple but useful one and the interested reader is referred to [7] and [1] for details; an overview of the model between rainfall and rainfall excess is now given.

In principle, rainfall excess could be calculated by estimating any 'losses' due to evapotranspiration and changes in catchment wetness, by employing measurements of hydrometeorological variables and soil moisture. However, this approach is usually data intensive and not necessarily accurate because of sampling, measurement and modelling errors. Also, it may not be necessary for streamflow modelling over short periods in humid temperate regions where the effects on streamflow of variations in evapotranspiration over restricted periods are usually of secondary importance to variations in rainfall.

The approach adopted in IHACRES is to account for variations in catchment wetness (i.e. the 'ripeness' of the catchment to produce streamflow at the time of the causative rainfall) by maintaining a running index of exponentially weighted past rainfall. A single optimal parameter which determines the length of 'memory' for exponentially weighting the past rainfall can be determined fairly rapidly by u er-

Rainfall excess is calculated by multiplying r_k^* by s_k at each time step and then scaling to ensure equality between volumes of rainfall excess and streamflow over the calibration period. Thus effective rainfall, u_k, is given by

$$u_k = const.\, r_k^* s_k \qquad (6)$$

where *const.* is the scaling factor.

Features and illustrative applications of IHACRES

1. Calculation of rainfall excess

The package allows the user to use the non-linear model (4) - (6) to calculate rainfall excess or for other calculated values of rainfall excess to be accepted directly.

Experience with the above simple non-linear model for rainfall excess ([7], [1], [3-6]) has shown it to work well in combination with (2) for a variety of catchment types and rainfall/evaporation regimes. Fig. 2 shows the rainfall and the IHACRES-derived rainfall excess over a 490 day period from 22 August 1976 for the Kenwyn at Truro, a 19 km² catchment in southwest England. Notice in Fig. 2 how, as the catchment wets up due to recent rainfall, rainfall excess approaches rainfall, after a relatively dry spell rainfall excess is only a small proportion of rainfall.

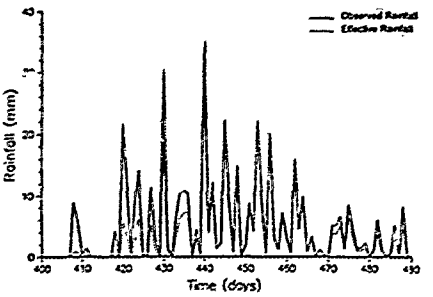


Fig. 2. Daily rainfall and rainfall excess for the Kenwyn at Truro from 22 August, 1976.

II. Calibration-validation of IHACRES models

Several methods of parameter estimation are available in IHACRES to calibrate catchment-scale models of streamflow behaviour in response to rainfall excess. These are basically various forms of instrumental variable estimation (see [1] and [8] for more details). When any model has been calibrated it should be validated using data from another period of record. Fig. 3 shows the calibration model-fit to daily data over a three year period from late July 1982 for the 894 km² Teifi at Glan Teifi in West Wales. This model has six parameters (four in the transfer function given by (2), one for the calculation of the catchment wetness index and one for adjusting rainfall for seasonal temperature-related evapotranspiration effects) and explains about 89% of the variance in the streamflow. To assess how well the model performs on another period of the Teifi record, Fig. 4 shows it applied to daily data for a three year period from the beginning of August 1976. Here, the model calibrated on data from 1982 to 1985 explains about 85 per cent of the initial variance of streamflow in the earlier period of record. The model captures the dynamics of the Teifi catchment hydrology and it is evident, therefore, that reasonably consistent unit hydrographs for tota streamflow and for quick and slow flow components have been identified.

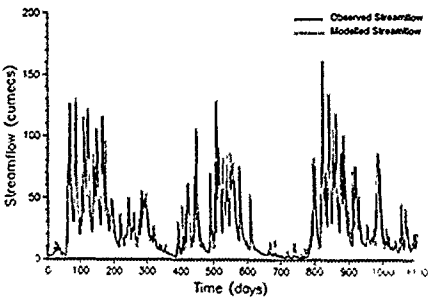


Fig. 3. Calibrated model fit to daily Teifi streamflow at Glan Teifi from 25 July 1982.

However, careful inspection of Figs 3 and 4 reveals that there are some systematic discrepancies between observed and modelled streamflow For example, many of the small hydrograph peaks in summer seasons are overestimated, indicating that the model could still be improved The fairly large difference between observed and modelled streamflow at the beginning of the wet season of 1986/87 (i.e. at the start of record shown in Fig. 4) corresponds to the 'wetting up' period at the end of the 1976 Drought and therefore presents a particularly demanding period of record to model. Littlewood and Jakeman [6] examine the 30 year record of Teifi daily streamflow and catchment areal rainfall to (a) improve the model-fit to summer events and (b) assess the temporal invariance of the unit hydrograph and its quick and slow components.

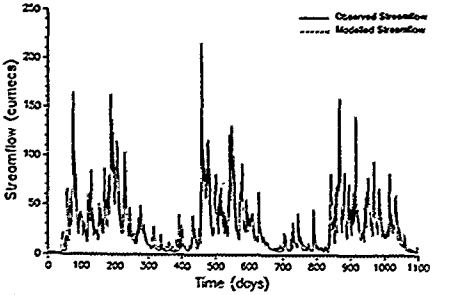


Fig. 4. Validation of model used for Fig. 3 on daily Teifi streamflow from 1 August 1976.

III. Separation of hydrographs into quick and slow flow

Separate convolutions of the quick and slow first-order transfer function components of (2) with rainfall excess generate the quick and slow flow components of streamflow respectively. Fig. 5 shows the modelled daily total streamflow and the corresponding slow flow component over the three year period from 1974 to 1976 for the 743 km² Colne at Denham, a southward draining tributary of the Thames in England. Much of the catchment comprises Chalk geology and the proportion of Colne streamflow which can be considered baseflow from a regional groundwater reservoir is therefore high. The high slow flow component of streamflow derived by IHACRES shown in Fig. 5 accords well with the provenance of a relatively large volume of baseflow from the Chalk

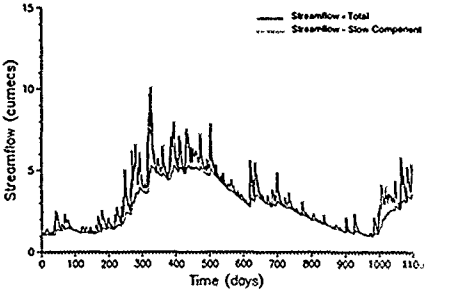


Fig. 5. Modelled daily Colne streamflow and slow flow at Denham from 1 January 1974.

In contrast, Fig. 6 shows the modelled hourly total streamflow and the IHACRES slow flow component over a 400 hour period from late July 1981 for a 0.01 km. catchment (Maimai, M6) in New Zealand. Here, it can be seen that the quick flow (total minus slow) is the dominant component during events, particularly at and near peak flows Interestingly, it has been estimated from isotope tracer field experiments that the provenance of most of the streamwater at and near peak flows in the M6 Maimai catchment is 'old' water which was in the catchment prior to the causative rainfall [9].

In order to explain a dominance of 'old' water in peak streamflows in such upland catchments, where slopes are steep and soils are thin, it is necessary to invoke a displacement mechanism of streamflow

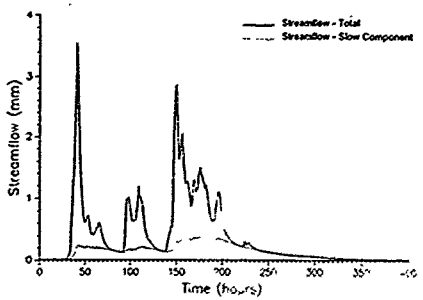


Fig. 6. Modelled daily Maimai (M6) streamflow and slow flow from 25 July 1981.

generation. Acceptance of both (a) the conclusions of the isotope tracer studies and (b) the IHACRES results indicates that the quick flow and 'old' water components have more in common in terms of provenance than the quick flow and 'new' water components. However, the underlying conceptualisation of streamflow generation in IHACRES is runoff from a catchment of variable wetness (or, alternatively, variable contributing area) represented by an idex based on an exponential weighting of past rainfall. Arguably, therefore, IHACRES is more in accord with widely accepted theory of streamflow generation than the theories of displacement necessary to explain the isotope tracer results. Furthermore, IHACRES hydrograph separation gives results which are intuitively reasonable for both high and low baseflow streams.

Clearly, there are differences of opinion in the scientific community about the basic nature of the dominant streamflow generation process in small upland catchments in humid regions. Application of IHACRES, using concurrent groundwater, rainfall and streamflow time series observations, and comparison with tracer techniques in particular field studies, may provide a way forward in any attempt to resolve these differences.

IV. Assessing and monitoring the impact of land-use change

If good models of the form given by (2) can be established from rainfall and streamflow data over different periods of interest, then the component flow characteristic catchment properties given by (3) can be derived for each of those periods. Furthermore, since the instrumental variable techniques used in IHACRES include computation of approximate confidence intervals for the parameters in (2), it is possible, by a Monte Carlo technique, to compute approximate uncertainties associated with the characteristics in (3) over any calibration period.

Table 1 shows how the slow flow time constant (τ_s) and relative volume throughput (V_s) for the 7.7 km^2 Monachyle catchment in Scotland change over a four year period (1984/85 - 1987/88) during which a part of the catchment was artificially drained before conifer trees were planted [4]. Each value shown in Table 1 was derived from a time series of daily rainfall and streamflow for annual periods starting in mid-June. While the Table shows large uncertainties on the characteristic properties (i.e. the ranges defined in parentheses), the mean values indicate that an increasing relative volume of flow passes through the slow flow storage following draining.

In another investigation [5] of the effects of land cover change on the hydrological response characteristics of catchments, IHACRES is being applied to examine rainfall and streamflow data from a 20 km^2 upland catchment in the ACT, which was instrumented following a substantially damaging bushfire in 1983. Analysis of data sets from 1983 until 1990,

TABLE I
MEAN VALUES FOR SOME CHARACTERISTIC CATCHMENT PROPERTIES FOR THE MONACHYLE CATCHMENT.

	1984/85	1985/87	1987/88
τ_s	42.98	39.63 (12, 68)	34.95 (12, 64)
V_s	0 05	0.15 (.08, 0.25)	0 22 (0.13, 0.31)

* Available approximate 70 per cent confidence intervals are shown in brackets

when revegetation had largely been re-established, reveals the stronger influence of evapotranspiration in the latter part of the data sets. Fig. 7 shows the unit hydrograph and its quick and slow components for this catchment, derived from model calibration over the two year period immediately following the bushfire.

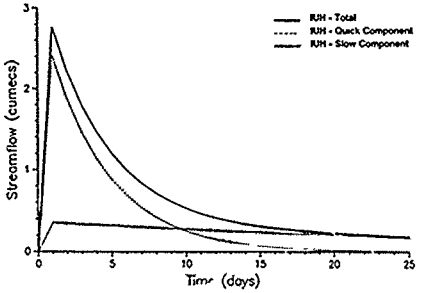


Fig 7 Unit hydrograph and compo ent flows for model calibrated from daily streamflow at Licking Hole Creek from 1 April 1983 to 13 March 1985.

Concluding remarks

The paper has outlined several features of IHACRES and has presented some of its applications (existing and potential) covering a range of problems in hydrological science and environmental management. The success to date of the methodology in terms of goodness-of-fit to measured streamflow, both in calibration and validation modes, and the ability to characterise well-defined quick and slow components of streamflow, augur well for application of IHACRES in areas of environmental and engineering hydrology quite generally. Attempts to improve the performance of IHACRES include (a) investigation of alternative rainfall excess models — to better model small runoff events in summer seasons and (b) utilisation of local groundwater or soil moisture content measurements — to account for variations in the slow flow component of streamflow.

Acknowledgements

Data for the Kenwyn, Teifi and Colne were provided from the UK Surface Water Archive databases maintained by the Institute of Hydrology The Balquhidder catchments are being monitored by the Institute of Hydrology. Data for Licking Hole Creek were provided by Australian Capital Territory Electricity and Water. The Maimai data were kindly provided by Dr A J Pearce, Director, Forestry Research Institute, Christchurch, New Zealand. We are also very grateful to Shelley Santoso for her careful preparation of this manuscript.

References

[1] A.J. Jakeman, I.G. Littlewood and P.G. Whitehead, 'Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments', *J. of Hydrology*, Vol. 117, pp. 275-300, 1990.

[2] V.T. Chow (ed.), *Handbook of Applied Hydrology*. New York: McGraw-Hill, 1964.

[3] A.J. Jakeman, I.G. Littlewood and P.G. Whitehead, 'Catchment-scale rainfall-runoff event modelling and dynamic hydrograph separation using time series analysis techniques', In: *Computer Modelling in the Environmental Sciences*, D.G. Farmer and M.J. Rycroft (eds.), IMA Conference Series, Oxford University Press, 1991.

[4] A.J. Jakeman, I.G. Littlewood and P.G. Whitehead, 'Detecting the effects of land-use changes on the quick and slow components of streamflow in the Balquhidder catchments', *J. of Hydrology*, (in press).

[5] A.J. Jakeman and I.G. Littlewood, 'Analysis of the effects of bushfire devastation and subsequent revegetation on the hydrologic response of a catchment subject to strong seasonal variations in temperature', in preparation.

[6] I.G. Littlewood and A.J. Jakeman, 'An assessment of the temporal invariance of quick and slow streamflow component unit hydrographs for the Afon Teifi over 30 years of record', in preparation.

[7] P.G. Whitehead, P.C. Young and G.M. Hornberger, 'A systems model of stream flow and water quality in the Bedford-Ouse River: I. Stream flow modelling', *Water Research*, Vol. 13, pp. 1159-1169, 1979.

[8] P.C. Young, *Recursive Estimation and Time-series Analysis - an Introduction*. New York: Springer, 1984.

[9] M.G. Sklash, M.K. Stewart and A.J. Pearce, 'Storm runoff generation in humid headwater catchments: 2. A case study of hillslope and low-order stream response', *Water Resources Research*, Vol. 22(8), pp 1273-1282, 1986.

NEW RESULTS IN TURBULENT DIFFUSION AND THEIR CONSEQUENCES FOR AIR QUALITY ASSESSMENT

P C CHATWIN
Department of Applied and Computational Mathematics
The University of Sheffield
Sheffield, S10 2TN, England

G W GOODALL
Department of Mathematics and Statistics
Brunel University, Uxbridge
Middlesex, UB8 3PH, England

N MOLE
Department of Mathematics
University of Essex, Wivenhoe Park
Colchester, CO4 3SQ, England

PAUL J SULLIVAN
Department of Applied Mathematics
The University of Western Ontario
London, Ontario, Canada, N6A 5B9

Abstract
It is increasingly recognized that, to be satisfactory, decisions involving air quality require the use of stochastic tools. This paper summarizes recent research findings and discusses the influence they should have on the development and interpretation of statistical models that will be used to support prediction, assessment, and monitoring procedures.

I INTRODUCTION

Historically, the first mathematical models used in air quality work assumed that the process of pollutant dispersion within the atmosphere was a deterministic process. Casual observation, for example of smoke from an industrial chimney or a cigarette, shows on the contrary that the process is stochastic (i.e. random). Figure 1 shows two time series of pollutant concentration in the atmosphere (chosen from many possible examples at random!). Both illustrate the inherent and large variability characteristic of all such series. This variability is ignored by deterministic models, such as Gaussian plume models and many other more complicated schemes, which, though readily available commercially (invariably linked up with eye-catching graphics packages!), have little connection with reality. Some countries still use such models in legislation or in quasi-legal regulations. It is to be hoped that they do not lead to disasters before they are replaced by scientifically sound models, and these must be stochastic.

The development of stochastic models for air (and water) quality assessment is now a thriving research activity - and quite rightly so. Important advances have been made by Jakeman and his colleagues at ANU, Canberra (see Jakeman, Bai Jun and Taylor 1988, and references therein). The present paper summarizes some recent work by the authors who predominantly approach the topic from the discipline of theoretical physics. In selecting material to include, preference has been given to work that is important in constructing stochastic models and not necessarily to research results of most fundamental scientific interest.

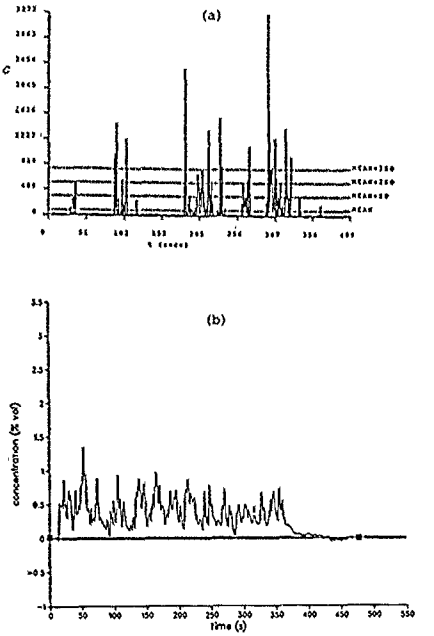


Figure 1: Typical concentration time series from field trials. (a) Mylne (1989), (b) Nielsen and Jensen (1991), Chatwin and Goodall (1991).

II THE PDF OF CONCENTRATION

Let Γ be the concentration of a species (foreign gas, particles, ...) dispersing in the atmosphere. As Figure 1 shows, Γ is a random function of position x and time t. Its probability density function (pdf) $p(\theta;x,t)$ is the most basic measure of variability. In the usual way, $p(\theta,x,t)$ is defined for $\theta \geq 0$ by

$$p(\theta;x,t) = \frac{d}{d\theta}\{prob[\Gamma(x,t) \leq \theta]\}, \qquad (1)$$

and the mean μ and variance σ^2 of Γ satisfy

$$\mu = \int_0^\infty \theta p d\theta, \quad \sigma^2 = \int_0^\infty \theta^2 p d\theta - \mu^2. \qquad (2)$$

In conventional turbulent diffusion work, μ and σ^2 are usually denoted by $\overline{c}$ and $\overline{(c')^2}$ respectively. Note that, in general, μ and σ^2 depend strongly on x and, for releases of pollutant that are not steady, on t. In view of widespread confusion and lack of precision, it must be emphasized that time-averages do not occur in the formulation above. Means are defined by equations like (2) and must be estimated from data by appropriate statistical methods which are not,

in general, time-averages. For environmental reasons there may be interest in the statistical properties not of the (instantaneous) concentration Γ but of Γ_T, its time-average over some defined interval T (eg 5 min or 1 hr). The pdf of Γ_T is not simply related to $p(\theta;x,t)$; in particular it must depend on T. Given also that, except for very large T, the pdf of Γ_T is more difficult to model δ, and, in general, to measure than that of Γ, it is timely to ask for hard scientific evidence (rather than an appeal to conventional practice, misguided legislation or even anecdotal wisdom) that Γ_T for a specified non-zero T is the appropriate measure of environmental danger, for example from toxic gases. A continued failure to consider this point seriously will inevitably result in wrong, even fatal, decisions being taken.

III FACTORS INFLUENCING PDF MODELLING

From the point of view of environmental decision support, it is important to seek robust and practical adequate mathematical models of $p(\theta;x,t)$; ideally such models ought to be related to the particular environmental danger (e g. flammability or toxicity?) and they should be simple. Jakeman and his colleagues, and others, have tested many candidates (lognormal, gamma, Weibull, ...) against data. Chatwin and Sullivan (1989) showed that $p(\theta;x,t)$ has the exact structure:

$$p(\theta,x,t) = \gamma f(\theta;x,t) + (1-\gamma)g(\theta;x,t), \qquad (3)$$

where $\gamma = \gamma(x,t)$ is an intermittency factor (equal to μ/θ_0, where θ_0 is the source concentration), and f and g are themselves pdfs. The importance of the formulation (3) is that $\gamma \ll 1$ except near the source so that $p \approx g$, i.e. the concentration statistics are dominated by the transfer (e.g. by molecular diffusion but excluding advection) of pollutant from air emanating from the source to air not emanating from the source. This observation has important consequences for modellers.

If simple models exist, they must enable parametrization to take account of atmospheric conditions and source geometry. In particular there would be great benefit to be derived if it could be shown that there was a simple relationship connecting quantities like μ and σ^2. The physically well-based equation

$$\sigma^2 = \beta\mu(\alpha\mu_0 - \mu), \qquad (4)$$

where α and β are constants and μ_0 is the maximum value of μ at a cross-section downwind of the source, is of this type, and (4) has been validated for many datasets by Chatwin and Sullivan (1990). However all these datasets were from steady releases in laboratory experiments; datasets from experiments more relevant to environmental hazard assessment are much rarer but work is in hand to see whether (4) can be generalized. Particular interest attaches to instantaneous or - more generally - unsteady releases.

Even though, in principle, $p(\theta;x,t)$ is determined by the governing equations of fluid dynamics and mass transfer, there is no method known (or likely to be discovered soon) of using these equations to obtain exact results about $p(\theta,x,t)$, or approximations of known accuracy to it. There is therefore high uncertainty about any proposed mathematical model and, consequently, an essential requirement before such models are used as decision support tools, namely that they are validated against relevant high-quality data. Given the fine spatial structure (down to scales of order 10^{-4}m) present in all pollution, the collection of such data is extremely difficult. As a consequence data is influenced by unavoidable instrument smoothing. This is illustrated in Figure 2. Mole - see Mole (1990) and Mole and Chatwin (1990) - has examined this problem and established its practical importance. He has also considered the different, though superficially related, topics of instrument (and other) noise and of thresholding.

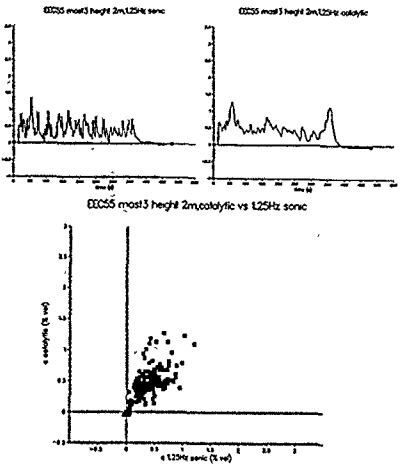


Figure 2: Top diagrams show time series obtained (for same concentration) by two different instruments. Bottom diagram is a scatter diagram comparing the results, the points are not close to a line of slope 1 through the origin - in fact $r \approx 0.7$ (Chatwin & Goodall 1991).

Another area of research of the same genre is the magnitude and removal of statistical noise arising, again inevitably, because of the limited size of datasets.

There is no space to consider the summarized problems in depth, but interested workers should consult the cited papers, and also Sullivan (1984) and Sullivan and Yip (1989).

IV ACKNOWLEDGEMENTS

For many years our research has been supported by agencies that include the Natural Sciences and Engineering Research Council, and the Department of National Defence (both of Canada), the Science and Engineering Research Council, and the Ministry of Defence (both of the UK), the Commission of the European Communities and NATO We are grateful to them.

REFERENCES

Chatwin P C & Goodall G W 1991 *Brunel University Department of Mathematics & Statistics Technical Report* TR/02/91.
Chatwin P C & Sullivan P J 1989 *Phys Fluids A1*, 761-763.
Chatwin P C & Sullivan P J 1990 *J Fluid Mech* 212, 533-556.
Jakeman A J, Bai Jun & Taylor J A 1988 *Atmos Envir* 22, 2013-2019
Mole N 1990 *Atmos Envir* 24A, 1313-1323
Mole N & Chatwin P C 1990 *Brunel University Department of Mathematics & Statistics Technical Report* TR/16/90
Mylne K R 1989 In *Proc 4th Int Workshop on Wind and Water Tunnel Modelling of Atmos Flow and Disp* (Karlsruhe 1988, ed A G Robins)
Nielsen, M & Jensen N O 1991 *Report No* Risø-M-2923 (Risø National Laboratory, DK-4000, Roskilde, Denmark)
Sullivan P J 1984 In *Proc 4th Joint Conf on Applics of Air Poll Meteor* (Portland, Oregon, ed A G Beals and N E Browne), 115-121
Sullivan P J & Yip H *Final Report on Contract No W7702-9-R114/01-XSG* (Def Res Estab Suffield, Ralston, Alberta, Canada TOJ 2NO)

Modeling and Simulation of Contaminant Transport and Sedimentation Processes in Fluvial Systems

D. Prochnow, H. Bungartz, Ch. Engelhardt
Institute of Geography and Geoecology
Rudower Chaussee 5
O-1199 Berlin
Germany

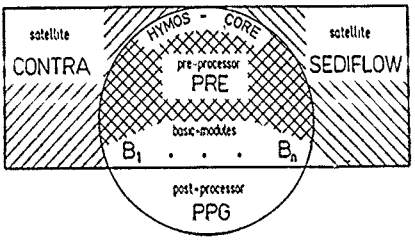
Abstract – In addition to experimental data, modeling and simulation of the flow and transport processes are important means to describe and predict the water quality in fluvial systems. With the help of a verified mathematical model one can investigate the principles and connections in an aquatic ecosystem by simulation of different process variants (e g. transport of dissolved and particulate substances, determination of particle load, self-purification by settling of particulate matter, exchange processes between bottom sediment and the water column, consequences of water management activities) The presented computer program HYMOS has been developed to handle these problems. Its structure, its capabilities and an example of its application are shown here.

I. Introduction

Rivers in industrial countries are often polluted with dissolved or particulate substances from anthropogenic sources, and a considerable amount of dissolved matter can be adsorbed by the suspended sediments carried in rivers. Hence the fate of particulate phases of nutrients and pollutants is governed by sediment transport. Adsorption-desorption transfer, particulate transport as well as deposition and erosion determine the path of a substance in aquatic systems [5,6].

There is a great need for powerful mathematical models predicting hydrophysical, chemical, and biological processes in fluvial systems. The adequate mathematical tools of these processes are coupled systems of nonlinear partial differential equations of hydrodynamics (Reynolds equations and special convection diffusion equations) connected with initial and boundary conditions [1,2]. Using the HYMOS computer code, one can solve typical initial and/or boundary problems based on the hydrological specified equations mentioned above.

II. The HYMOS program package



CONTRA simulates groundwater problems
SEDIFLOW simulates surface water problems

Fig. 1 Program package structure

The HYMOS program package consists of a core and of several satellites using this core. One can divide the core into a pre-processor (named PRE), into some basic modules, needed by all satellites, and into a post-processor (named PPG), see Fig. 1 The numerical method to solve initial and boundary value problems for partial differential equations implemented in the core of HYMOS is a special finite element technique, called the multi-bases Galerkin finite difference method [4] Until now two satellites of HYMOS exist which are named CONTRA and SEDIFLOW CONTRA deals with groundwater problems Moreover, this program is able to identify automatically parameters of considered flow and transport processes in porous media using observation data. The satellite program SEDIFLOW is concerned with contaminant transport and sedimentation processes in rivers and lakes. In the further presentation essential properties of SEDIFLOW are considered in more detail.

III. Capabilities of the SEDIFLOW computer code

Using SEDIFLOW one can simulate flow, contaminant transport, exchange and transformation processes as the following.

- turbulent flow in real river segments
- settling and transport behavior of particle fractions (e g continuous side discharge or after accident)
- transport of dissolved and particulate contaminants and nutrients and their accumulation in bed sediment
- complex exchange and transformation processes of substances, like volatility on the free surface, decay of substances in the dissolved or particulate phase or in the sediment, sorption effects concerning particle fractions or the sediment, aggregation and particle growth
- coupled transport and exchange processes of several substances

The user of the SEDIFLOW code can choose between three models of turbulent momentum exchange.

- constant coefficients of turbulent momentum exchange
- vertical dependent exchange coefficients (mixing length model)
- the k-ε-turbulence model

SEDIFLOW manages two- or three-dimensional problems. All provided solutions vary with time. Steady state solutions are found as final states of time dependent processes.

SEDIFLOW also allows to apply reduced models. Reducing a mathematical model is a common way to get first rough information about an ecosystem as well as to limit computational costs [3]. Following reduced models are implemented in SEDIFLOW

- two-dimensional vertical cross sections (2-D version)
- depth-averaged models (shallow water situations)
- static pressure assumption (instead of full momentum equations)

The application of SEDIFLOW results in

- flow characteristics :
  - velocity fields of the mean turbulent flow
  - distribution of the isotropic pressure (water level)
  - quantities, describing the local state of turbulence (turbulence kinetic energy, dissipation rate of turbulence energy, production rate, coefficients of turbulent momentum exchange)
- characteristics of substances
  - concentrations of an unlimited number of suspended particle fractions, dissolved substances and particle-sorbed or sediment-accumulated substances

- secondary quantities as particle deposition and entrainment rates, net deposition rates, sediment accumulation rates, mean substance concentrations in the water body

IV. SEDIFLOW application example

In Fig s 2 and 3 are shown the predicted flow and the settling behavior of different particle fractions along a slow flow river section. In the settling basin after a sudden increase of the discharge cross section, the discharge velocity decreases from $v_E = 0.3$ m/s at the inflow to downstream values of about 0.02 m/s (Fig. 2).

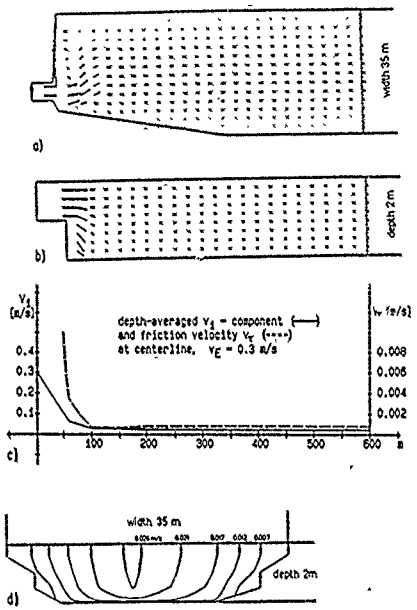


Fig. 2 Settling basin, flow characteristics
a) calculated horizontal velocity field near free surface
b) vertical cross section, located at inflow centerline
c) depth-averaged longitudinal velocity and bottom friction velocity
d) distribution of longitudinal velocity at downstream cross section

The total particle mass of 9 mg/l measured upstream is separated into four different fractions classified by the particle settling velocity (1st fraction - smallest settling velocity). Fig. 3a shows the individual fraction settlement along the considered river section From this result it is evident that the very light matter, corresponding to the first fraction, is eroded because of the overcritical values of friction velocity at the inflow area (Fig 2c) Fig 3b compares the fractional particle settling with the behavior of the total particle mass using a median value of settling velocity.

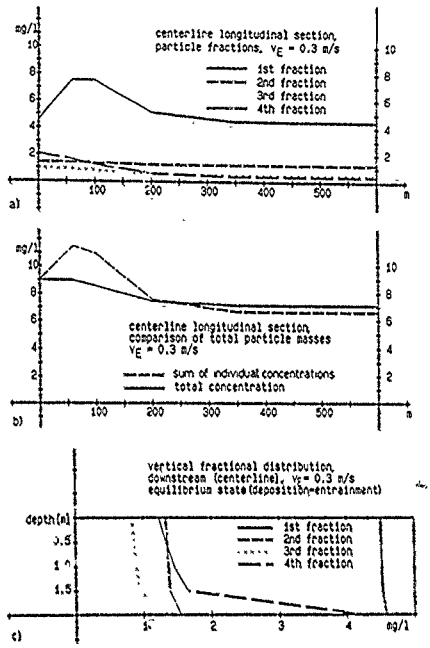


Fig. 3 Settling behavior of individual particle fractions

References

1 *ASCE Task Commitee on turbulence models in hydraulic computations. Turbulence modeling of surface water flow and transport.* J Hydraulic Engineering, 114 (1988)9, 970-1073

2 King,I.P *Strategies for finite element modeling three-dimensional hydrodynamic systems* , Advances in Water Recources, 8 (1985), 69-76

3 O'Connor,D.J , *Models of Sorptive Toxic Substances in Freshwater Systems I - Basic Equations* , J of Environ. Engineering, 114 (1988)3, 507-532

4 Prochnow, D. *On the Galerkin Finite Difference Method: A Multi-Bases Approach.* Int. J. Num. Meth. in Engin , 21 (1985), 713-723

5 Prochnow, D , Burgartz, H , Engelhardt, Ch ,*Modeling Suspended Sediment Transport in Fluvial Systems Including Retardation.* Syst. Anal Model. Simul , 8 (1991)6, (in press)

6 Westrich, B. *Flussaler Feststofftransport - Auswirkungen auf die Morphologie und Bedeutung für die Gewässergüte* Schriftenreihe gwf Wasser und Abwasser , R Oldenbourg Verl , München Wien, 1988

Testing Models for High Concentration Solute Transport in Porous Media by Analytical Solution of a Simple Flow Problem

Michael Thiele
Institute of Geography and Geoecology
Rudower Chaussee 5
O-1199 Berlin F.R.G.

Abstract - The classical forms of Darcy's and Fick's laws seem to be inadequate to describe miscible solute transport through porous media at higher density and viscosity gradients. In addition to numerical computer codes accompanying laboratory experiments to identify possible model extensions, it is proposed to use asymptotic expansion solutions of the governing equations for a two-dimensional laminar stable mixing layer flow. The method is applied to a non-linear extension of Fick's law [1] which is proved experimentally, up to now, for a one-dimensional flow configuration [2]. Thus there is a need for reexamining the existing experimental data [3],[4] and for some new experiments.

I. Governing Equations

The basic equations describing fluid flow and solute transport in a porous medium are the equations of conservation of total and solute mass. These are, for a conservative solute in a source-free flow field [1]

$$\frac{\partial(n\rho)}{\partial t} + \nabla(\rho \mathbf{v}) = 0 \tag{1}$$

and

$$n\rho\frac{\partial(c/\rho)}{\partial t} + \rho \mathbf{v}\nabla(c/\rho) + \nabla \mathbf{J} = 0 \tag{2}$$

where n is the medium porosity, ρ is the fluid mass density, $\mathbf{v}$ is the fluid velocity vector, ∇ is the Nabla operator, c is the solute volume concentration $[M/L^3]$, $\mathbf{J}$ is the solute dispersive mass flux $[M/(L^2T)]$ and t is time. These equations must be supplemented by relations of conservation of momentum. In most problems of slow incompressible fluid flow in porous media, it is assumed that the velocity and pressure P are related by Darcy's law

$$\mathbf{v} = -\frac{k}{\mu}(\nabla P - \rho \mathbf{g}) \tag{3}$$

where μ and k are, respectively, the fluid's dynamic viscosity $[M/(LT)]$ and the porous medium's permeability $[L^2]$ which is assumed to be isotropic and homogeneous, $\mathbf{g}$ is the gravity acceleration vector. A commonly used expression for the dispersive mass flux is

$$\mathbf{J} = -\rho \mathbf{D}(\mathbf{v})\nabla(c/\rho) \tag{4}$$

where $\mathbf{D}$ is the Bear-Scheidegger dispersion tensor [5]. It contains in an additive manner the porous medium attenuated molecular diffusion of the solute and the mechanical dispersion, the last being proportional to the mainstream flow velocity over a wide range of interesting parameters. As proportionality constants serve two characteristic quantities, the longitudinal α_L and the transversal α_T dispersivities $[L]$ of the isotropic porous medium. To close the system constitutive relations are needed for the fluid's density and viscosity which are their functions of the solute concentration c. Linear relationships between c and ρ or μ are assumed to be valid here, though the real dependences are at least slightly non-linear for most water soluble solids

$$\rho = \rho_0\left(1 + \delta\frac{c}{c_s}\right) \quad \text{and} \quad \mu = \mu_0\left(1 + \varepsilon\delta\frac{c}{c_s}\right) \tag{5}$$

with c_s - the maximum solute concentration occuring in a specified transport problem and $\delta = (\rho_s - \rho_0)/\rho_0$, $\rho_s = \rho(c_s)$ and $\varepsilon\delta = (\mu_s - \mu_0)/\mu_0$, $\mu_s = \mu(c_s)$, where ρ_0 and μ_0 are, respectively, the density and viscosity values at zero or some datum concentration.

In recent years a growing community of experimentators has been collecting material evidencing the inconsistency of either Darcy's or Fick's law or both to describe the solute transport at high concentration in the case of natrium chloride ($\delta_{max} = 0.2$, $(\varepsilon\delta)_{max} \approx 1.0$) [2]–[4], [6], [7] Hassanizadeh [2] has tested seven extensions of Fick's law (4) with the help of a discretisized numerical model of the governing equations on a one-dimensional experimental set-up. The breakthrough behaviour of saltwater at the outlet of a column initially filled with fresh water was measured and compared with the calculated curves corresponding to the different models. Only the following extension of Fick's law met all the experimental values satisfactorily

$$(1 + \beta|\mathbf{J}|)\mathbf{J} = -\rho \mathbf{D}(\mathbf{v})\nabla(c/\rho) \tag{6}$$

where β is a new constant which could be identified uniquely. To avoid difficulties of the mentioned numerical procedures and to proceed further with the verification of the non-linear solute mass flux ansatz (6) it is proposed to deal with a mixing layer flow problem as described below.

II. The mixing layer flow problem

Consider the parallel stationary flow of saltwater ($z < 0$) and freshwater ($z > 0$) which start to mix at $x = 0$, see Fig. 1. To ensure almost horizontal flow the inlet velocities of both fluids have to be related to each other as depicted in Fig. 1, [3]. In [3], [4] steady-state distributions of the diluted salt were measured along the vertical at some distance L from the inlet edge which was large compared to the width of the transition zone from salt water to fresh water. In [3] there was provided an approximate analytical solution to this mixing problem based on the classical forms of Darcy's (3) and Fick's (4) laws, too. Within the frames of (3) and (4) the proposed dependence of the transversal dispersivity α_T on c, could not account for the measured distributions either. It would have implied, however, a challenge to the geometrical concept of dispersivity that should be independent of the fluid flowing through the porous medium.

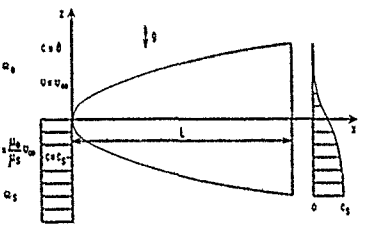


Fig. 1 Freshwater/saltwater mixing layer flow with transition zone

III. Asymptotic expansion for large values of x

Without mixing the pressure distribution P of the introduced flow problem would be, [3],

$$z > 0: \qquad P = P(0,0) - \frac{\mu_0 u_\infty}{k} x - \rho_0 g z \tag{7}$$

$$z < 0: \qquad P = P(0,0) - \frac{\mu_0 u_\infty}{k} x - \rho_i g z \tag{8}$$

i.e. a sharp interface between the two fluids would be retained. Mixing causes perturbation of this pressure distribution which further will be denoted by p.

In the theory of fluid motions it rendered possible to get a partial insight into the structure of the solutions of certain flow configurations by the method of asymptotic coordinate-type expansions [8]. This method is most suitable for finding the solution of the flow and transport problem at large distances from the interaction zone of the body and the fluid instead of in the whole flow region, or, as in our case, at large distances from the inlet edge $x = 0$. In analogy to [8], it can be shown that the perturbation pressure p and the dimensionless concentration $\Theta = c/c_i$ can be expanded in a power series of $x^{-1/2}$ for large values of x in the following way:

$$p(x,z) = p_{-1}(\eta)\, x^{1/2} + p_0(\eta)\, x^0 + p_1(\eta)\, x^{-1/2} + \dots \tag{9}$$

$$\Theta(x,z) = \Theta_0(\eta)\, x^0 + \Theta_1(\eta)\, x^{-1/2} + \dots \tag{10}$$

where $\eta = z/\sqrt{4\alpha_T x}$ is the so called similarity variable. Substitution of these sums into the system (1)-(3),(5),(6) leads to ordinary differential equations for each of the coefficient functions p_i, Θ_i. These differential equations are successively solvable and the solutions uniquely determinable only for some first coefficient functions, and logarithmic terms in x have to be added to the series (9) and (10), [9]. However, the terms of practical interest of the Θ-expansion, Θ_0 and Θ_1, could be determined after having linearized the respective equations by means of an additional δ-power series expansion. Up to linear terms in $x^{-1/2}$ and δ the concentration distribution reads:

$$\begin{aligned}\Theta \;=\; & \frac{1}{2}\left(1 - \operatorname{erf}\frac{\eta}{2}\right) + \frac{\epsilon\delta}{4}\left(\operatorname{erf}^2\frac{\eta}{2} - 1\right) - \\ & - \frac{\sqrt{\alpha_T k g \rho_0 \delta}}{4\sqrt{\pi}\mu_0 u_\infty} \exp\left(-\frac{\eta^2}{4}\right) \operatorname{erf}\frac{\eta}{2}\, x^{-1/2} - \\ & - x^{-1/2}\, \beta \frac{c_i u_\infty \sqrt{\alpha_T}}{4\sqrt{\pi}} \exp\left(-\frac{\eta^2}{4}\right) \operatorname{erf}\frac{\eta}{2} \cdot \\ & \left[1 + \frac{\delta}{4}\left(\operatorname{erf}\frac{\eta}{2} - 2\right) - \frac{\epsilon\delta}{4}\left(5 \operatorname{erf}\frac{\eta}{2} + 2\right)\right]\end{aligned} \tag{11}$$

The first three terms are identical with the solution obtained in [3]. In Fig. 2 there is shown the influence of the additional β-term resulting from the flux ansatz (6) for a fixed set of remaining parameters taken from an experiment in [4]. Unfortunately this term cannot be fitted with the experimental distribution for no value of β either, although some basic trends of the influence of the entering parameters are identical with those observed. Nevertheless, it is believed that the utilization of the asymptotic expansion method to solve the flow problem considered is an appropriate means to 'visualize' the effects of transport model extensions like (6) and others on measurable quantities in a straight forward manner. Further research endeavour both experimental and analytical should be undertaken. It is planned to study the influence of several extensions of Darcy's law [10],[11] and its combination with the Fick's law extension (6).

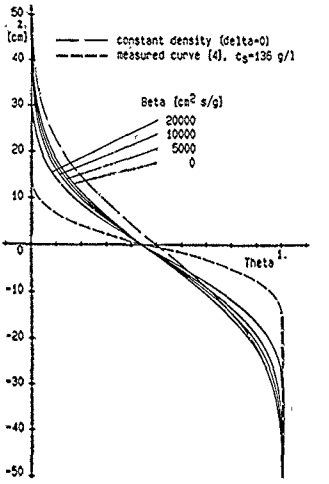


Fig. 2 Influence of the β-term on the vertical concentration distribution at $L = 1115$ cm (parameter values are taken from [4]: $\alpha_T = 0.135$ cm, $u_\infty = 0.0414$ cm/sec, $\delta = 0.088$, $\epsilon\delta = 0.255$, Peclet number $Pe = 1450$)

1 Hassanizadeh, S.M., *Verification and validation of coupled flow and transport models.* GEOVAL-90 Symposium, May 1990, Swedish Nuclear Waste Inspectorate, SKI, Stockholm, 1990

2 Hassanizadeh, S.M., ModelCARE-90, Proc. Conf. The Hague, Sept. 1990, IAHS Publ. no. 195, 241-250, 1990

3 Rinnert, B., Mitt. Inst. Wasserbau Nr. 52, Univ. Stuttgart, 1983

4 Spitz, K., Mitt. Inst. Wasserbau Nr. 60, Univ. Stuttgart, 1985

5 Bear, J., *Dynamics of fluids in porous media.* American Elsevier, New York etc., 1972

6 Kobus, H.E., K. Spitz, Proc. 21st IAHR Congress, Melbourne, August 1985, 170-174, 1985

7 Bues, M.A., L. Zilhox, J. Hydrology, 120, 125-141, 1990

8 Chang, I-Dee., J. Math. Mech., 10, 811-876, 1961

9 Stewartson, K., J. Math. Phys., 36, 173-191, 1957

10 Herbert, A.W., C.P. Jackson, D.A. Lever., Water Resour. Res., 24, 1781-1795, 1988

11 Hassanizadeh, S.M., Adv. Water Resour., 9, 196-222, 1986

WATER ENTRY SIMULATION OF BLUNT BODIES ON BOUNDARY FITTED COORDINATE

N. NISHIKAWA, and TOSHI. ABE
Faculty of Engineering, Chiba University
Yayoi 1-33 Chiba, Japan

Abstract In this paper finite difference calculations for water-entry cavity are performed for the cylinder, circular disk, and hemisphere. One of the employed finite difference schemes is SOLA-VOF scheme on Cartesian grid and other two schemes are on BFC(Boundary Fitted Coordinate)grid with partly or fully contravariant velocity expression. The symmetricity of the flow is assumed. An time dependent incompressible Navier Stokes equation is treated. Monitoring display of the contours of flow variables is also attempted nearly simultaneously with the fluid dynamic calculation on a graphic workstation.

I. INTRODUCTION

As for water entry cavity flow on a small scale, the splash phenomena around sphere with centimeter size are experimentally studied by Kuwabera et al[1], and with numerical approach, such flows were attacked[2,3] by MAC schemes assuming the flow is axisymmetric.

In first part of this paper the water entry cavity flow is numerically analyzed with the boundary fitted coordinates(BFC). Initially the lower vertex of body is tangentially contact with the upward liquid flow and a body is fixed at initial position. The Navier stokes equation and continuty equation are transformed to the forms on the generated BFC grids. HSMAC method is modified for the transformed forms, whose results are compared with those on the cartesian grid.

Following the authors' successfully animated scenes of the splash of a droplet[4], but in contrast to pre or post-processing, a target of the present study is the attempt to visualize the flow behavior simultaneously with the fluid dynamical calculation.

II. BASIC EQUATIONS

The symmetry of flow and incompressibility are assumed. An incompressible Navier Stokes equation is written for a general curviline ar coordinate as follows

$$\frac{\partial V}{\partial t} + \nabla\cdot(VV) = -\frac{1}{\rho}\nabla p + \nu\Delta V + g \quad (1)$$

$$\nabla\cdot V = 0 \quad (2)$$

Physical domain is transformed to a rectangular domain of general coordinate ξ, η. Equations (1),(2) are nondimensionalized by $\rho U^2/R$, where U,R are the oncoming fluid velocity and body radius, respectively.

A. Numerical Scheme

It has been a important subject how tactically a staggered differencing is applied on non-cartesian grids without oscillatory pressure or velocity. The reasons of the oscillation is that the transformed velocity component are not aligned with coordinate directions[5]. First, here, the non-physical velocity formulation is applied. Finite differencing of the convection terms employes a third order upwind difference expression Except this the scheme is the modification of HSMAC scheme to a general coordinate.

B. Relaxation for continuity equation

Time integration of Navier Stokes equation (1) by explicit Euler method gives the renewed velocities at a new time level. Pressure and renewed velocities are adjusted to satisfy the continuity equation(2) whose residual is expressed with D=div V before convergence. That is, the relaxation process is iteration between the pressure expression which is given by the diagonal dominant approximation and the new estimate for the velocities as follows

$$\delta p = -\omega D/\Delta t(g_{11}+g_{22}) \quad (3)$$

$$\delta u = -\Delta t(\xi_x\delta p_\xi + \eta_x\delta p_\eta)$$
$$\delta v = -\Delta t(\xi_y\delta p_\xi + \eta_y\delta p_\eta) \quad (4)$$

where ω is the iteration factor and g_{11}, g_{22} are the metric coefficient e.g. $g_{11}=\xi_x^2+\xi_y^2$, and subsctipts denotes the derivatives.

C. Boundary Condition

The markers without substantial mass are shifted at each time step with interpolated-velocity from velocioties at the cell sides. Then the location of free surface is simply determined from the location of surface(SUR) cell beyond which no cell involve the marker.

In the results on BFC shown here, the surface tension is not included so far.

The gas pressure in the open cavity can be calculated by the Bernoulli equation. While, if a closed cavity appears, the adiabatic process is applied for the pressure inside the bubble behind the obstacle.

Thus, the atmospheric gas does not interact with liquid flow except in the closed cavity. In accordance with this, we applied the initial flat free surface condition.

D. Numerical Results by BFC scheme

Numerical results are shown for Re=30000, 40000, to compare with the existing experiment for Re=38720(R=8mm, U=2.42m/s if ν=0.01 cm^2/s). The velocity vectors are shown in Fig. 1 for Re=30000. It is natural that the large velocity appears at the peak near the body. In contrast in Fig.3 at large velocity, Re=40000, such film cannot be predicetd.

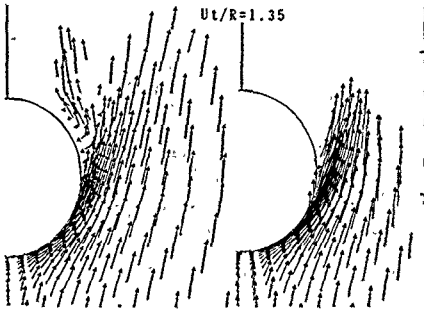


Fig.1 Re=30000

Fig.2 Re=40000

The free surface profiles at typical time-levels are shown in Fig.1. The two peaks appears in the developed state.The peak near the body is the leading edge of the liquid filt along the body surface.Such liquid film is known in the experiment[1]. The height of the peak is rather lower than the experimental value where leading edge reaches 75 degree from lower stagnation point when Ut/D=1/2 .

While, the physical component formulation [5] have been coded and sucessfully applied to a cylinder submerged in a uniform flow by the younger author Mr.Abe.This code is expected to produce an improved behavior.

III CARTESIAN GRID and SOLA-VOF SCHEME

We introduce here the results[3]on cylindrical coordinate r,z, respectively,correspond to radial and axial coordinate,and for which equations(1,2) are rewritten

In this case axisymmetric obstacle such as hemisphere is treated upon the same assumption as in the BFC formulation. As in usual MAC scheme a staggered differencing is applied.The finite difference approximation employs a two-point upwind scheme.

A. Free surface condition

Near the free surface we employ VOF[volume of fluid] function,which represents the liquid volume fraction. The volume fraction F satisfies the following equation

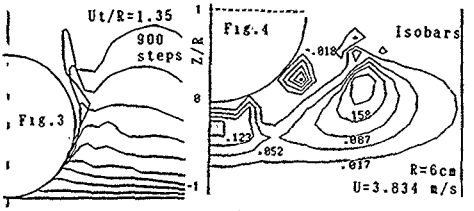
$$\frac{\partial F}{\partial t} + \frac{1}{r}\frac{\partial (rFu)}{\partial r} + \frac{\partial Fv}{\partial z} = 0 \quad (5)$$

The liquid-surface-pressure measured from gas phase can be expressed with surface tension σ by radii of curvatures.

As described in the previous section II for the BFC formulation,the pressure and renewed velocities are adjusted to satisfy the continuity equation

B. Results on Cartesian Grid

We introduce the numerical examples which



is an additional results to the authors' previous study[3]. The numerical examples are given for the hemisphere with diameter 12cm and the uniformflow velocity:U= 3.834m/s . As in the aut'ors previous report,our results[3] show qualitatively good agreement with the results by SMAC calculation[2].

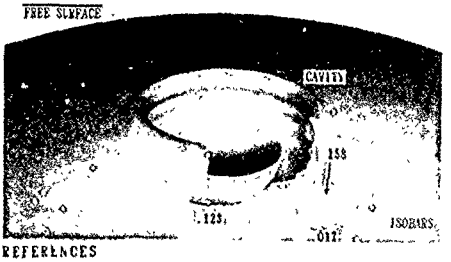
In Fig.4 the isobars are shown for dimensionless time t=Ut/D=3,834. Even in the surface profile,we cannot find sharp peak climing upward. That is the liquid film cannot be predicted by the cartesian mesh code as far as with the cell sizes $\Delta r = \Delta z = R/20$.

The isobars as in Fig.4 or the equivorticity lines can be displayed during calculation. On our weorkstation TIATN-I the free surface shape in the surface modeler image as in Fig.5 can be displayed nearly in real time mode. That is,the CPU-time are about 20 seconds,10 sec respectively for CFD operation and isobars displaying. The appropriate intervalfor display is estimated with the comparison of the elapse time for contour display with that for the CFD calculation.

IV CONCLUSIVE REMARKS

Finite differnce scheme on the BFC grid predict liquid film along curved surface ,which could not be predicted by the scheme on the Cartesian grid.Monitoring display scenes on the workstation is shown as the proposal of a visual aid for code development process.

Fig.5 Free Surface and Isobars



REFERENCES

[1] Kuwabara,G.,Tanba,H.,and Kono,K.,J.Phys. Soc. Japan,Vol.56(1987),pp2733-2743.

[2] Chen,J. & Yan K. Acta Aerodynamic Sinica, Vol.4,p47(1986)

[3] Nishikawa.N,Y.Kito,and T.Abe, Computer & Fluids(or Japan-USSR Sympo. CFD p217(1990).

[4] Nishikawa,N. Suzuki,Akiyama.S.,Proc.10th ICNMFD p499-504(1986).

[5] Demirdzic,I et al.,Computer & Fluids 15, p 251(1987).

IDENTIFICATION AND CLASSIFICATION OF POLLUTING SOURCES IN CLOSED BASINS

V.G.DOVI'
ISTIC- Universita' di Genova
Via Opera Pia 15
16145 Genova (ITALY)

AND

Z. SALIMOV
UsSSR Academy of Sciences
Gorki Street 77
707100 Tashkent (USSR)

Abstract- Identification of source strenghts of closed basins is often difficult due to the presence of diffuse sources or the unreliability of official information. A source - receptor technique is developed in this communication for the estimation of polluting sources at the boundary of the basin from concentration measurements inside the basin.

I. INTRODUCTION

The identification of polluting sources in closed basins has become more and more important in the last decade. Contaminations due to industrial wastes and municipal sewage waters are presently threatening the survival of biological life in many lakes and interior seas or can give rise to dangerous eutrophication phenomena or become centers of infections. Hence the necessity of monitoring water quality, of identifying polluting sources and of classifying them. Unfortunately the identification problem can be carried out using receptor techniques only. In other words the location and the nature of polluting sources at the boundary of the basin are to be determined using the information provided by concentration measurements inside the basin. This is due to the fact that sources can be unknown such as diffuse sources, generally connected with cumulating microcontaminations, or their official strengths unreliable, as in the case of more than one country having access to the basin. To tackle the problem we first have to setup a model connecting concentrations in the basin with fluxes of contaminants at the boundary; then a suitable algorithm has to be constructed for the model inversion, i.e. the determination of fluxes using experimental values of the concentration.

II. MODEL SETUP

The diffusion of pollutants in water has been assumed to obey the following fickian law:

$$\epsilon\left(\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}\right)=r \qquad (1)$$

In other words diffusion has been supposed to be prevailingly horizontal, whereas a more general dispersion law r, including deposition, vaporization and chemical reactivity is assumed to hold vertically. This term is strictly connected with the so called water purification power and , as a first approximation, it can be assumed constant. Thus the overall problem is defined by

$$\Delta^2 c=\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}=k=r/\epsilon \qquad (2)$$

$c(x_i,y_i)$ known within experimental accuracy at location $(x_i,y_i), i=1,\dots,N$

$$-\epsilon\frac{\partial c}{\partial n}=\Gamma(S)$$

where Γ is the unknown pollutant flux along the boundary contour S (n being the normal direction to S).

Since $\Gamma(S)$ is unknown, the Neumann problem can be replaced by the equivalent Dirichlet problem

$$\Delta^2 c=k \qquad (3)$$
$$c(S)=\gamma(S)$$
$$c_i \text{ measured at } (x_i,y_i)$$

Once $\gamma(S)$ has been estimated, it is straightforward to estimate $\Gamma(S)$.

III. FORMAL SOLUTION

The solution of equation (3) can be written as $c(x,y)=c_1(x,y)+c_2(x,y)$

$$\Delta^2 c_1=k \qquad \Delta^2 c_2=0$$
$$c_1(S)=0 \qquad c_2(S)=\gamma(S)$$

Formally

$$c_1(x_0,y_0)=K\int G_I(x_0,y_0|x,y)\,dx\,dy$$

where G_I is the Green's function of the first type of the non-homogeneous Laplace equation with homogeneous Dirichlet's conditions.

If the conformal mapping

$$f:S\to(\xi\mid\ \|\xi\|^2),\ z_0=(x_0,y_0)\to(0,0)$$

(i.e. f maps S onto the unit disk with x_0, y_0 mapping into the origin , $\xi=f(x_0,y_0|x,y)=f(z_0,z)$) is known, we can compute G_I from the relation [1]

$$G_I(x_0y_0|x,y)=\frac{1}{2\pi}\ln\frac{1}{|f(z_0,z)|}$$

Similarly c_2 can be computed using the Poisson integral

$$u(r_0\phi_0)=\frac{1}{2\pi}\int_0^{2\pi}\frac{1-r_0^2}{1+r_0^2-2r_0\cos(\phi-\phi_0)}\gamma(\phi)d\phi$$

$r_0\le 1;\qquad 0\le\phi\le 2\pi$

In fact if the conformal mapping defined above is known, we can write $\gamma(\phi)=\gamma(f(S))$ and compute $u(r_0\phi_0)$ from the Poisson integral and eventually obtain

$$c_2(x,y)=u[f^{-1}(r_0,\phi_0)]$$

Thus in general we can write the formal solution of equation (3) as

$$c(x,y)=k\int G_I(x,y|\xi,\eta)d\xi d\eta+\int G_{II}(x,y|S)\gamma(S)dS$$

where both G_I and G_{II} are known if the conformal mapping $f:G\rightarrow\{z\mid \|z\|^2\leq 1\}$ is known.

Setting

$$d(x,y)=c(x,y)-k\int G_I(x,y|\xi,\eta)d\xi d\eta \qquad (5)$$

we obtain

$$d(x_i,y_i)=\int G_{II}(x,y|S)\gamma(S)dS \qquad (6)$$

IV. ALGORITHMIC DEVELOPMENT

There are now two major problems to be addressed. First a procedure was to be developed for the numerical computation of the conformal mapping and consequently of the d_i and G_{II} as defined in equation (6).
Secondly equation (6) has to be solved. Being a Fredholm integral equation of the first type it is a badly posed problem in the sense of Hadamard and needs regularization before being inverted [3].

A. Computation of the conformal mapping

We shall use an extremum property of conformal mapping to carry out its numerical evaluation (see [4] for the demonstration).
Let us consider the set of mappings
$H_p=\{f \mid f \text{ continuous in } G, \|f\|<\infty,\ f(p)=0,\ f'(p)=1\}$
where the norm $\|.\|$ is defined as

$$\|f\|=\max_{z\in G}|f(z)|=\max_{z\in\partial G}|f(z)|$$

and the norm $|.|$ is arbitrary.
∂G indicates the closure of G . It can be proved that there exists a mapping $\hat{f}$ such that
$r=\|\hat{f}\|<=\|f\| \qquad \forall f\in H_p$

This mapping is the conformal mapping

$$\hat{f}:G\rightarrow D(r)=\{\zeta \mid |\zeta|\leq r\}$$

If $|f(z)|=(Re^2[f(z)]+Im^2[f(z)])$
it can be proved that r=1 and $\hat{f}$ is the sought for conformal mapping
$\hat{f}:G\rightarrow\{z \mid \|z\|^2\leq 1\}$
A suitable approximation to $\hat{f}$ is provided by

$$\hat{f}(z)=w_1+\sum_{j=2}^{n}a_jw_j$$

with $w_j=(z-p)^j$ and the coefficients a_j can be computed from

$$\left\{Re^2[f(z_k)]+Im^2[f(z_k)]\leq\eta\right\}\quad \eta=\min \qquad (7)$$

We can either choose a high value of n or divide ∂G into a suitable number of elements ∂G_m such that $\partial G=\sum\partial G_m$ and use in each of them a low polynomial approximation to $\hat{f}$.

Problem (7) is a general minimization problem subject to nonlinear constraints. Though high dimensional, it can be solved without insurmountable convergence difficulties using either augmented lagrangian or successive quadratic programming techniques ([5] , [6]).

B. Regularization techniques

Equation (6) is an ill-posed problem in the sense of Hadamard, i.e. arbitrarily small changes in the data can give rise to considerable deviations in the estimation of $\gamma(S)$. Hence the necessity of regularizing the problem , i.e. of picking among all the solutions $\gamma(S)$ which fit the data d_i within experimental accuracy, the one which fits additional a-priori information best. In the literature the additional information is generally connected with some extremum criterion such as the smallest, the smoothest, the most randomly distributed, etc. There is no apparent reason for applying similar criteria in this case. Instead an approach based on fuzzy set theory is employed. This approach, which has been described in detail elsewhere [7], makes it possible to use all the qualitative information available to regularize the problem and obtain the solution that, to an assigned level of confidence, is consistent with both experimental data and a-priori information.
Only a concise outline is given here. The additional fuzzy information available is written in the form

$$f_i\ [B_i(\gamma)]$$

where B_i indicates some qualitative and / or quantitative property of the expected solution γ. $B_i:\Gamma\rightarrow K$ [where K is a suitable space for the set of properties B_i. Thus for instance K might be the set (strong source, moderate source, negligible source)] and
$f_i : K\rightarrow L$
where $L=\{x\mid x\in\Re, 0\leq x\leq 1\}$
In other words f_i is a subjective membership function which assigns a numerical value in the range 0 - 1 to the elements of the set K.
More complex formulations are necessary if both f and B are more complicated.
The case in which f is characterized by four parameters has already been examined by the Author and a complete flow-chart of the corresponding algorithm has been described in detail [7].
Thus the overall identification algorithm can be written as

$$\|A\underline{\gamma}-\underline{d}\|^2=\min$$

subject to

$$f_{i,\min}\leq f_i[B_i(\underline{\gamma})]\leq f_{i,\max}$$

where $f_{i,\min}$ and $f_{i,\max}$ are values included in the range (0,1) and depend on our degree of confidence of the solution having certain properties. Thus the theoretical and numerical structure has been completely outlined.

V. CONCLUSIONS

A method for monitoring coastal contributions to the overall pollution using concentration measurements inside closed basins has been described. Simulated data and convenient fuzzy constraints have been used to validate the data. More insight will be gained from the use of real data with the possible introduction of general fuzzy bounds. This work has been started and is presently underway in our institutes.

REFERENCES

[1] Sveshnikov A. and A. Tikhonov, The theory of functions of complex variables, Mir Publishers, Moscow (1971), p.193

[2] Ref. [1] p. 191

[3] Tikhonov A. and V. Arsenine, Methodes de Resolution de Problemes Mal Poses, Mir Editors, Moscow (1974)

[4] Opfer G., New Extremal Properties for Constructing Conformal Mappings, Numer. Math., 32, 423-429, (1979)

[5] Hestenes M.R. , Multiplier and gradient methods, J. Optim. Theory Applns., 4, 303-320 (1969)

[6] Locke M.H., R. Edahl and A.W. Westerberg, An improved successive quadratic programming optimization algorithm for engineering design problems, AIChE Jl. 29, 871-874, (1983)

[7] Dovi' V.G. - Use of fuzzy optimization algorithms for the solution of a class of ill-posed problems in data analysis, Computers Chem. Engng., 14, 9, 957-966 (1990)

CONTROL SYSTEM DESIGN IN THE ENVIRONMENT

Marko Šega
J.Stefan Institute, 61000 Ljubljana, Jamova 39, Yugoslavia

Abstract- In the work a view to control system design for real application is presented. The way, the design is carried out, and proposed control system (hopefully also realized) are naturally tightly dependent on known vagueness, e.g. compromises between "exact" and experimental approaches, linear and nonlinear treatment, etc. The decisions, which are mainly in designer's hands, hang not only on control theory and designer's expertise (knowledge and abilities) but also on technical, economic and sociologic characteristics of the environment, in which and/or for which control system is designed. In opposition to control system design the functionality of simulation tools, which are in application examples the most frequently used in all phases of control design, is much more independent from the environment and designer. In the work described aspects are discussed through modeling, analysis and control system design for one of several subprocesses in the factory which produces titanium dioxide.

I. INTRODUCTION

The principal activity at the Department of Computer Automation and Control is devoted to research, development and applications in the area of automatic (computer-based) control of systems and processes. This orientation expresses the intention to cultivate and expand knowledge in the control field, but much more the needs and the possibilities in our community. Through the applications we try to contribute to the development, i.e. we solve concrete problems, show and stress the role of control systems as well as we help in forming staff for industry departments.

In the paper modeling, analysis and control system design for one of 19 subprocesses in the factory which produces titanium dioxide are presented. Although the full automation of this production is new (as I am informed) my intention is not to go in details, but to discuss the environmental aspects in relation to control design methodology and simulation tools. While the methodology is rather specific and subjective, I believe, that described functionality for simulation tools is the minimum for successful work.

II. PROCESS DESCRIPTION

The factory, where the production of titanium dioxide is based on the hydrolysis of sulfate solution /1/, was built approximately 15 years ago. But lately, the world trends towards automation which ensures better supervision and higher product quality convinced progressive staff in the factory to invest in the research. Our department was engaged and till now we have implemented a control system for removal of ferrous sulfate by crystallization. The next subprocess we are studying and which is discussed in this work, is the reduction of the ilmenite solution.

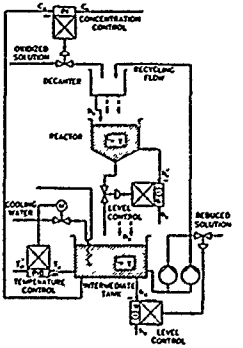


Fig. 1. Simplified Technological Scheme of the Subprocess

The goal of the reduction is to reduce all the ferric components of the ilmenite solution to the ferrous state and to convert a small proportion of the titanium dioxide to the trivalent condition. The goal is reached by the reaction with metallic iron. The reaction is exotermic and mass-transfer limited. So a specific technology is used, where three parallel reactors with (semi)batch character (metallic iron should be added and washed every 8 hours; every 8 days a reactor should be emptied, washed and refilled) are connected and managed in a way which ensures continuous production. The simplified technological scheme of the reduction is shown in Fig.1.

From the input/output point of view the reduction in Fig.1 represents a kind of continuous stirred tank reactor. The decanter is included only to spread the input and the recycling flow in all active reactors. The intermediate tank gathers the outputs from the reactors and enables heat removing with only one cooling system. The reduced solution is taken from the intermediate tank.

III. MODELING

The model base was taken from the literature /2/, where also typical assumptions and simplifications which are suitable for the initial stage are described, e.g. concentrated parameters, perfectly mixed solution, constant density, first order reaction. With the additional simplifications, e.g. a cylindric form (with a spare effective base) of the reactors and unique reaction (the input solution reacts irreversibly to form the reduced solution with desired concentrations) which takes place only in the reactors, first mathematical model was derived. It consists of three differential equations for each reactor, three differential equations for the intermediate tank, one differential equation for the cooling system and one algebraic equation for the decanter.

As usually the first mathematical model did not agree enough with the measurements. It had to be improved, but corresponding physical and chemical background was unknown. The project was time and funds limited, so there were also very small chances to get detailed information about the process, i.e. to repeat the measurements with better equipment and experiments. So the modeling process relied only on engineers reasoning and experimental approaches.

With the time the equations of the first model were improved with elements which are related to additional nine differential equations, that can be reasonably argumented (e.g. two consecutive reactions and two "parazite" reactions slower simplified model reaction, delays because of the transport and mixing, heat transmission to metallic iron and vessels, the influence of the measurement equipment) This engineering and experimental phase in modeling was much more difficult and lasting especially because the final model consists of 21 parameters (6 known, 7 estimated from the available data and measurements and 8 unknown).

In my opinion the modeling process was not very specific for our environment. For the initial stage of control system design, for projects with smaller control requirements or for time and funds limited projects I find such an approach quite natural. Obviously also differences can occur. As the designer I see some differences even in our environment. They are mainly conditioned with the extent measurement equipment, the possibilities to experiment on the object and understanding the needs for experimental results. Finally they can occur also because of more complex reasons, like economic status and valuating of the expert work.

IV. CONTROL DESIGN

The whole control design was influenced by the environment. At its beginning there were two main reasons. First, the economics dictated not to invest too much in the equipment what was a severe limitation already on the measurement equipment. Therefore, any technologic changes were out of question, although the original factory plan differs from the realized one (in the plan each reactor has its own decanter, input and recycling flow). Secondly, the factory staff was rather content with existent control. While at the same time, it was also burdened too much, it was not enough acquainted with the reduction and often

did not think enough in the control sense. Thus, the identification of control goals was a hard job and could only be done through few interviews, intensive model behaviour studies and engineer reasoning. So, after relatively long period three global control goals were derived, i.e. reduced solution should have desired concentration, the speed of production should be controlled and the temperature in the reactors ("reaction temperatures") should be constant.

From the control goals it can be easily realized that the existent control scheme is not quite appropriate. Namely, only the first global control goal can be satisfactorily fulfilled. The other two goals are partly fulfilled by operators (set point for the level control in the reactors; influences the production speed) or by technologists (set point for the temperature control in the intermediate tank; "indirect control" of the reaction temperature). Naturally, such man-in-the-loop control demands quite a big skill and, in any case slower and less precise control.

In the first design step the control scheme was simply improved by introducing a speed production controller (on the basis of the production flow it sets the set point for the level control in the reactors) and reaction temperature controller (on the basis of the reaction temperature it corrects the set point for the temperature control in the intermediate tank). This approach seemed to be very attractive because it reaches the main goals. However, such control structure is not good enough because it "neglects" the conditions in the reactors (although the temperature in the intermediate tank and desired concentration of the product are reached the concentrations in the reactors and reaction temperatures can quite deviate). This is especially significant by the concentrations, while the reaction temperatures are tightly connected to them. So modified control structure, which is shown in Fig.2, was derived. Its main characteristics are:

- The concentration control is realized for each reactor separately.
- The production speed control is not needed in the stacionary point (when the desired concentration is reached) because the input and output flows are the same.
- Due to the fact that output flow is not a function of the concentration in the intermediate tank but function of the level in this tank, a proportional controller for quality assurance is included (it corrects the production speed when the desired concentration is not reached; it should be dominant over the production speed in nonstacionary situations).
- The reaction temperature control is the same as in the first design step.

Simulation results show that all global control goals as well as the desired concentration in each reactor can be satisfactorily reached. Indeed, some problems with the reaction temperature remain (there exist three controlled variables and only one control input, i.e. the temperature in the intermediate tank), but they are less critical because the concentrations in the reactors are under control. If the modified control structure will proof itself in the reality, the only arguments against it could be of the environmental character, i.e. problems with the acceptance (e.g. familiarity and the necessity for handling two different, i.e. computer and hardware control concept).

V. SIMULATION TOOL

For the modeling and control design procedure our simulation tool SIMCOS /3/ was used. It is a CSSL based simulation language which is permanently improved with some nonstandard options, like features that make the language closer to hybrid simulation, features that enable simulation of digital control systems and simulation in real time.

With no regard to the interactiveness (which is not so important for experienced users) a good simulation tool for control system design should be as functional and flexible as equation oriented simulation languages are. Besides this fundamental demands, which are fulfilled in several simulation tools, some others elements are necessary. In the field of methods such elements are real time simulation (including possible link with the "foreign" objects, i.e. with the process for verification of control strategy or with the computer control system for verification of control system), optimization, curve fitting, linearization and parameter studies. In the field of functionality which belongs to "infrastructure" the necessary elements are hierarchical blocks, accessible data base for and from other CACSD tools (besides time histories, transfer function and state space description also the topological data base) and an opened (as much as possible) architecture to include specific models, methods and experiments.

I find proposed functionality in simulation tools independent from the environment and examples. Namely, almost all described functional elements would be quite useful in solving the discussed and other examples I made, and I believe also in any other control system design.

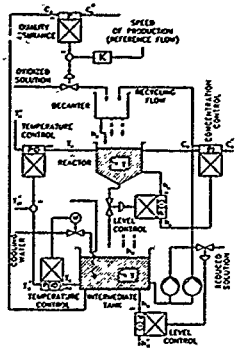


Fig. 2. Modified Control Structure for the Reduction

VI. CONCLUSIONS

In the work a control system design procedure for the reduction of the ilmenite solution is presented. It can be interesting for the community which is envolved in the automation of the titanium dioxide production. However, the main stress of the work is given to the environmental aspects that influence the control design procedure.

In the Department of Computer Automation and Control we try to cultivate, educate and expand control knowledge in our environment where too small attention was devoted to control problems in the past. We realize this intention mainly through concrete applications where not only the gap between theory and praxis should be solved but also several environmental effects should be considered. The status in the industry, like approaches to project management, funds and control knowledge, has quite a big influence on the control systems. This fact can also be confirmed by several control systems which have been implemented in our factories. Namely, the systems often do not function any more (the desired behaviour can not be reached or they are even put out of the function) or they are not quite actual nowadays (like in the example where original control structure does not consider the production speed as it was probably not so important in the past).

The paper gives also a view to the functionality of simulation tools used in control system design. Described simulation capabilities are quite independent from the environment and concrete problem and can be understood as a hint to tool designers as well as to potential users.

REFERENCES

/1/ Barksdale, J. (1966). TITANIUM - Its Occurrence, Chemistry and Technology. The Ronald Press Company, New York.

/2/ Luyben, W.L. (1973). Processs Modeling, Simulation, and Control for Chemical Engineers. McGraw-Hill Book Company, New York.

/3/ Zupančič, B , D Matko, R. Karba and M Šega (1987) SIMCOS digital simulation language with hybrid capabilities. Proc 4th Symp. Simulationstechnik, Zürich, pp 205-212

APPLICATIONS OF GAME THEORY IN LEGAL PROCEDURE OF COURT

Prof.N.K.HIPPARKAR
B.P.Sulakhe Comm.College, BARSHI,
Dist.Solapur, Maharashtra,(INDIA).

ABSTRACT : Legal Procedure of Court is shown as a game theory problem with Plaintiff (Complainant) and Defendant (Accused) as Players. Different witnesses, courses of actions, their moves etc. are taken as the strategies for these players. By defining the variable 'aij', we get the pay-off matrix. We solve the game by using any one of the methods. And on the value of the game, the judgement can be passed accordingly.

I. INTRODUCTION

A game is simply a set of descriptive rules. A play of the game includes every particular instance in which the game is played from begining to end. The participants are the players. A game consists of a sequence of moves of the players, while a play comprises a sequence of choices made by them. The decisive step in the mathematical treatment of games is the normalization achieved by introduction of pure strategies. A pure strategy is a plan formulated by a player prior to the start of a play, which covers all of the possible decisions which he may face during any play permitted by the rules of the game. The expected course of a play is thereby completely determined by the selection of a pure strategy by each player in ignorance of that chosen by any other player.

Zero sum two person games are played by two players and what one player wins, the other loses. A zero sum two person game with a finite number of pure strategies can be described by an mxn matrix $A = (a_{ij})$, where a_{ij} is the payment to player I by player II if I chooses his i^{th} pure strategy and II chooses his j^{th} pure strategy. Instead, during repeated plays of the game, he can be better off by mixing his various pure strategies randomly, but with chosen relative frequencies. Subsequently, a passage from such a 'sample' of plays of the game to the underlying population reveals the notion of a mixed strategy (mixed actions) as a probability distribution over the set of pure strategies. Mixed strategies for I and II will be denoted by x and y respectively, where

$x = (x_1, x_2, \dots x_m)$, $y = (y_1, y_2, \dots y_n)$;

$x_i \geq 0$ for all i and $y_j \geq 0$ for all j

with $\sum x_i = 1 = \sum y_j$

When I plays a mixed strategy x and II plays a mixed strategy y, the expected payment by II to I is G(x,y).

All discussion of zero sum two person games starts from the Minimax theorem which asserts:

$$\max_x \min_y G(x,y) = \min_y \max_x G(x,y)$$

The unique minimax value of G is called the value of the game A and will be denoted by v(A).

Mixed strategies x^* and y^* such that

$G(x^*,y) \geq G(x,y^*)$ for all mixed strategies x and y are called optimal or good strategies for player I and player II respectively.

II. LEGAL PROCEDURE : A GAME THEORY PROBLEM

II.1 INTRODUCTION : In legal procedure of court we can make use of 'Game Theory'. By this approach the legal advisors and magistrates may have sound and mathematical base in the course of formulating the judgement. Sometimes judgement is on discretionary powers of judges and it is possibility of confusion for them. But, I hope, if we apply game theory approach for court procedure, judges may have easier and sound base in passing the judgement. Also legal advisor may give more stress on certain point to prove the issues which are more important for his client benefit.

In court procedure, we have -

1. Two parties which are treated as players; Plaintiff (P) and Defendant (D).
2. Both these parties have conflicting interests.
3. Each party has a finite set of moves i.e. finite set of strategies.
4. An issue in favour of P is considered as an issue against D and vice-versa.

Here we assume player P as a maximising player and player D as a minimising player; thus player P wishes to $\max_x \min_y G(x,y)$ and player D wishes to $\min_y \max_x G(x,y)$.

II.2 NOTATION AND DEFINITIONS :

Let strategy sets for P and D be denoted by x and y respectively. We define the utility-matrix as follows.

$$A = P \begin{array}{c} \\ 1 \\ 2 \\ i \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & & j & & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & & & \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & & & \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \end{array} \quad D$$

where

a_{ij} : is the utility for player P when P chooses his i^{th} strategy and D selects j^{th} strategy.

($i = 1,2,\dots m$; $j = 1,2,\dots n$)

m, n are possible strategies for player P and for player D respectively.

Conventional game theory problems are either to search for a saddle point or use of mixed strategy method for selecting different strategies for different players. In real world problems however, the entries in the pay-off matrix are linguistically and qualitatively defined. In legal procedure the case is the same. In these type of situations it is necessary to model the problem with an imprecise pay-off table. In legal procedure context, to construct the utility matrix, one may take help of the previous records and experiences.

As a particular case, here we redefine the variable a_{ij} as follows

$a_{ij} = 1$, if an issue is passed by the magistrate in favour of P, when P adopts i^{th} strategy and D adopts j^{th} strategy;

$a_{ij} = 0$, if an issue is crossed i.e. effect of i^{th} strategy of P is nullified by the j^{th} strategy of D;

and $a_{ij} = -1$ if an issue is passed against P, i.e. passed in favour of D.

In this manner, we can obtain a utility matrix and it is treated as a pay-off matrix as in a usual game problem.

After obtaining pay-off matrix, we can solve the game by using any one of the methods. Let v(A) be the value of game. This value, v(A) will be very useful for magistrate in coming to judgements. Magistrate may have the following different alternatives depending upon the value of v(A)

i) If $v(A) > 0$, the judgement is passed in favour of P i.e. plaintiff wins the case;

ii) If $v(A) = 0$, the decision depends upon the discretionary power of judge and

iii) If $v(A) < 0$, judgement is passed against the P and it will be in favour of D i.e. Defendant wins the case.

III. REMARKS

Here it is seen that the pay-off matrix A has only three types of elements viz. $-1, 0$ and $+1$. It implies that all strategies have equal effect upon the decision of magistrate. But it is seen that different strategies may have different effects upon decision, e.g. some evidence, witnesses, situations, acts, etc. are more powerful than the other. To take into account this fact, we may use mixed strategies. Different strategies may be selected with different probabilities. If a particular strategy is more powerful in relation to decision making, then one may assign higher probability corresponding to it.

This approach is also very helpful to legal advisor in proving his client's side. He can find out the more important strategies for his client and can give more stress on the same to prove his point. Also, it is assumed that strategy sets x and y for players P and D are finite which is obvious otherwise there will not be an end to the case.

IV. CASE STUDY

For an illustration, let us consider the case. The reference is as follows : Reg. Civil suit No. 126 of 1981. In the court of Mr.M.H.Baig,

B.Sc.(Hons) LL.B.II

Joint Civil Judge.

J.D.Barshi at Barshi ½ Ex.No.41.

Dattatraya Nagnath Chikhale, Age 55 Plaintiff

Against

Navanath Rama Kumbhar, Age 40 years Defendant

In this case, Plaintiff and Defendant had following different strategies.

Plaintiff's Strategies :

Sr.No.	Description of Strategy
1.	Ownership of suit site
2.	Possession of suit site
3.	Ownership in tree
4.	Raise of obstrution by defendant
5.	Permission from municipal corporation to cut down the tree.

Defendant's Strategies :

Sr.No.	Description of Strategy
1.	Tenancy
2.	Right or interest in tree on suit site
3.	Possession of suit site
4.	Plaintiff obtained the permission to cut down the tree by misleading the corporation.

With these strategies and opinions of Hon.Judge, we obtain the following pay-off matrix.

		Defendant			
		1	2	3	4
Plaintiff	1	1	1	0	1
	2	1	1	0	1
	3	1	1	1	1
	4	1	1	0	1
	5	1	1	1	1

Solution of the problem :

We solve the game, whose pay of matrix is given as above.

		Defendant				
		1	2	3	4	Row min.
Plaintiff	1	1	1	0	1	(0)
	2	1	1	0	1	(0)
	3	1	1	1	1	(1)
	4	1	1	0	1	(0)
	5	1	1	1	1	(1)
column max.		(1)	(1)	(1)	(1)	

Here max. (Row min) $= \max_i \min_j (a_{ij}) = 1$

and min (column max.) $= \min_j \max_i (a_{ij}) = 1$.

Which implies

$\max_i \min_j (a_{ij}) = \min_j \max_i (a_{ij}) = 1$

Hence, the game has pure value $v(A) = 1$.

And, it is seen that $v(A) > 0$

Hence, as per our criteria the judgement is passed in favour of plaintiff i.e. in favour of Mr.Dattatraya Nagnath Chikhale and then decree is passed accordingly.

V. REFERENCES

1. Black-well, D(1965). Discounted Dynamic Programming, Ann.Math.Stat., 36, 226-235.
2. Filar J.A. and Raghavan, T.E.(1988): Algorithms of stochastic games, mimco.
3. Klein, E and Thompson, A.G.(1984): Theory of corrospondences, including applications to mathematical economics, Wiley-Interscience Publication, John Wiley and sons.
4. Maitra, A and Parthasarathy, T(1970): On Stochastic games, J.O.T.A., 5, 289-300
5. Parthasarathy T. and Singha, S(1989); Existence of Stationary equilibrium strategies in non zero sum discounted stochastic games with uncountable state space and state independent transitions, Int, Jour., Game Theory, 18, 189-194.

*****

Fortran Codes for Computing the Discrete Helmholtz Integral Operators

S. M. Kirkup

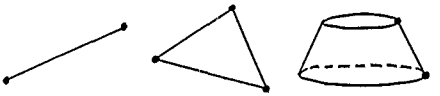
Address. Department of Mathematics and Computer Science, University of Salford, Salford, UK.

Abstract. In this paper Fortran Subroutines for computing the discrete form of the Helmholtz Integral Operators for two-dimensional, three-dimensional and three-dimensional axisymmetric problems are described. The subroutines are useful in the solution of Helmholtz problems via boundary element and related methods.

Introduction The Fortran subroutines described in this paper are useful in the implementation of integral equation methods for the solution of the general two-dimensional, the general three-dimensional and the axisymmetric three-dimensional Helmholtz equation

$$\nabla^2\varphi(\mathrm{p}) + k^2\varphi(\mathrm{p}) = 0 \qquad (1)$$

which governs φ in a given domain and k is a complex number termed the wavenumber. The subroutines compute the discrete form of the integral operators L_k, M_k, M_k^t and N_k that arise through the application of collocation to integral equation reformulations of the Helmholtz equation. Expressions for the discrete integral operators are derived by approximating the boundaries by the most simple elements for each of the three cases. The elements are illustrated in the figure.



Straight line element Triangular element Conical element

The subroutines are named HO2LC, HO3LC and HO3ALC and the parameters to the subroutine take the following form:

SUBROUTINE HO 2/3/3A LC (
Wavenumber (generally complex),
Point (p *and the unit normal* $\mathrm{n_p}$ *associated with* p),
Geometry of element (vertices which define element and unit normal),
Quadrature rule (abscissae and weights for computing integral),
Choice of discrete form required (L_k, M_k, M_k^t and/or N_k),
Answers (the values of the chosen discrete forms))

The Helmholtz Integral Operators. The Helmholtz integral operators are denoted L_k, M_k, M_k^t and N_k and they are defined as follows:

$$\{L_k\mu\}_\Gamma(\mathrm{p}) \equiv \int_\Gamma G_k(\mathrm{p},\mathrm{q})\,\mu(\mathrm{q})\,dS_\mathrm{q}, \qquad (2)$$

$$\{M_k\mu\}_\Gamma(\mathrm{p}) \equiv \int_\Gamma \frac{\partial G_k}{\partial \mathrm{n_q}}(\mathrm{p},\mathrm{q})\,\mu(\mathrm{q})\,dS_\mathrm{q}, \qquad (3)$$

$$\{M_k^t\mu\}_\Gamma(\mathrm{p}) \equiv \frac{\partial}{\partial \mathrm{n_p}}\int_\Gamma G_k(\mathrm{p},\mathrm{q})\,\mu(\mathrm{q})\,dS_\mathrm{q}, \qquad (4)$$

$$\{N_k\mu\}_\Gamma(\mathrm{p}) \equiv \frac{\partial}{\partial \mathrm{n_p}}\int_\Gamma \frac{\partial G_k}{\partial \mathrm{n_q}}(\mathrm{p},\mathrm{q})\,\mu(\mathrm{q})\,dS_\mathrm{q}, \qquad (5)$$

where Γ is a surface, $\mathrm{n_q}$ is the unit outward normal to the boundary at q and $\mu(\mathrm{q})$ is a bounded function defined for $\mathrm{q} \in S$. The vector $\mathrm{n_p}$ is a unit normal associated with p. $G_k(\mathrm{p},\mathrm{q})$ is the free-space Green's function for the Helmholtz equation: $G_k(\mathrm{p},\mathrm{q}) = \frac{i}{4}H_0^{(1)}(kr)$ in two dimensions and $G_k(\mathrm{p},\mathrm{q}) = \frac{1}{4\pi}\frac{e^{ikr}}{r}$ in three dimensions where $r = |\mathrm{r}|$, $\mathrm{r} = \mathrm{p} - \mathrm{q}$ and i is the unit imaginary number.

Subroutines HO2LC, HO3LC, HO3ALC. The form of the three Fortran subroutines are listed in this section The quantities that the subroutines compute are determined by replacing Γ by the appropriate element, μ by the unit function and G by the appropriate Green's function in (2)-(5). Results from the application of these subroutines to test problems are given in the reference.

```
*****************************************************************
*          Subroutine HO2LC by Stephen Kirkup   Dec 1990        *
*****************************************************************
This subroutine computes the discrete form of the 2-dimensional
Helmholtz integral operators LK,KK,Mkt, and Nk. Hence the subroutine
is useful in boundary element-type methods for the solution of Helmholtz
problems.
The subroutine has the form:

    SUBROUTINE HO2LC(K,P,NORMP,QA,QB,NORMQ,QPONEL,MAXNQ,NQ,AQ,WQ,
   * QLK,QNK,QNKT,QNK,DISLK,DISNK,DISNKT,DISNK)
    REAL*8 EPS
    INTEGER IU,LIMNQ
    LOGICAL CHECK
    PARAMETER (EPS=     ,IU= ,CHECK=    )

The PARAMETER statement
=======================
real EPS, integer IU, integer LIMNQ, logical CHECK.

The parameters to the subroutine
================================
complex K, real P(2), real NORMP(2), real QA(2), real QB(2),
real NORMQ(2), logical QPONEL, integer MAXNQ, integer NQ,
real AQ(MAXNQ), real WQ(MAXNQ), logical QLK, logical QNK,
logical QNKT, logical QNK, real DISLK, real DISNK, real DISNKT,
real DISNK.

*****************************************************************
```

```
*****************************************************************
*          Subroutine HO3LC by Stephen Kirkup   Dec 1990        *
*****************************************************************
This subroutine computes the discrete form of the 3-dimensional
Helmholtz integral operators LK,MK,Mkt, and Nk. Hence the
subroutine is useful in boundary element-type methods for the
solution of Helmholtz problems.

The subroutine has the form:

    SUBROUTINE HO3LC(K,P,NORMP,QA,QB,QC,NORMQ,QPONEL,MAXNQ,NQ,
   * XQ,YQ,WQ,QLK,QNK,QNKT,QNK,DISLK,DISNK,DISNKT,DISNK)

    INTEGER IU
    REAL*8 EPS
    LOGICAL CHECK
    PARAMETER (IU= ,EPS=     ,CHECK=     )

The PARAMETER statement
=======================
real EPS, integer IU, logical CHECK.

The parameters to the subroutine
================================
complex K, real P(3), real NORMP(3), real QA(3), real QB(3),
real QC(3), real NORMQ(3), logical QPONEL, integer MAXNQ, real NQ,
real XQ(MAXNQ), real YQ(MAXNQ), logical QLK, logical QNK,
logical QNKT, logical QNK, real DISLK, real DISNK, real DISNKT,
real DISNK.

*****************************************************************
```

```
*****************************************************************
*          Subroutine HO3ALC by Stephen Kirkup.   Dec 1990      *
*****************************************************************
This subroutine computes the discrete form of the 3-dimensional
axisymetric Helmholtz integral operators LK,Mk,Mkt, and Nk  Hence the
subroutine is useful in boundary element-type methods for the
solution of Helmholtz problems.

The subroutine has the form:

    SUBROUTINE HO3ALC(K,P,NORMP,QA,QB,NORMQ,QPONEL,
   * MAXNGQ,NGQ,AGQ,WGQ,MAXNTQ,NTQ,ATQ,WTQ,
   * QLK,QNK,QNKT,QNK,DISLK,DISNK,DISNKT,DISNK)

    PARAMETER (EPS=     ,IU= ,LIMNGQ=   ,LIMNTQ=   ,CHECK=     )

The PARAMETER statement
=======================
real EPS, integer IU, integer LIMNGQ, integer LIMNTQ, logical CHECK

The parameters to the subroutine
================================
complex K, real P(2), real NORMP(2), real QA(2), real QB(2),
real NORMQ(2), logical QPONEL, integer MAXNGQ, integer NGQ,
real AGQ(MAXNGQ), real WGQ(MAXNGQ), integer MAXNTQ, integer NTQ,
real ATQ(MAXNTQ), real WTQ(MAXNTQ), logical QLK, logical QNK,
logical QNKT, logical QNK, real DISLK, real DISNK, real DISNKT,
real DISNK.

*****************************************************************
```

Acknowledgement. The author is grateful to his colleague Dr. S. Amini for his advice on this work. The author is sponsored by a SERC - Admiralty Research Establishment grant.

Remark. The subroutines in this report are available through written request to the author.

Reference. S. M Kirkup (1990). *Fortran Codes for Computing the Discrete Helm-holtz Integral Operators.* Report MCS-90 09, Department of Mathematics and Computer Science, University of Salford, UK.

STABILITY ANALYSIS IN AERONAUTICAL INDUSTRIES

Stéphane Godet-Thobie
CERFACS
42, Avenue Gustave Coriolis
31057 Toulouse Cedex

Abstract - In this paper, we consider a large-scale non-symmetric problem that occurs in structural mechanics combined with aerodynamics ; we present in particular the Arnoldi method used in conjunction with a Chebyshev iteration technique, for computing a few eigenvalues of a large real non-symmetric matrix (ordered by their imaginary parts) and their associated eigenvectors. The work we report has been conducted at CERFACS in cooperation with AEROSPATIALE Aircraft Division and IBM France whose supports are gratefully acknowledged.

Introduction

Most of the large-scale nonsymmetric eigenvalue problems that arise in various research or engineering fields like mechanics or aerodynamics are related to the stability analysis of a physical system. At present, the matrices that appear in the largest calculations arise through the study of evolutionary problems like the Navier-Stokes equations or the differential systems arising in structural dynamics. As an example, the modelling of the behaviour of a plane in flight leads to a differential equation in time, which is then discretised by a finite element method. The computation of the modes of this physical system yields a nonsymmetric eigenvalue problem of the form .

$$Ax = \lambda x, \text{ where } A \in \mathcal{M}_{n\times n}(\mathbb{R}),\ \lambda \in \mathbb{C} \text{ and } x \in \mathbb{C}^n,$$

where the order of the matrix A can reach 5×10^5.

1 Why can nonsymmetric problems be difficult ?

The matrices arising in stability studies are often non-normal. A large departure from normality $\|AA^* - A^*A\|_F$ leads to an ill-conditioned eigenbasis which is always extremely difficult to compute. The matrix under consideration may also have multiple or defective eigenvalues. Such eigenvalues are not necessary ill-conditioned, but the calculation of their eigenbasis requires special methods like the block-Arnoldi or the Newton method [1]. When a few eigenvalues and eigenvectors of a large nonsymmetric matrix A are required, one can use a projection method on a Krylov subspace like the Arnoldi method or the nonsymmetric Lanczos method. The spectrum of the reduced problem is an approximation to a part of the spectrum of the original problem. matrix, which is interesting from the point of view of storage. The Arnoldi method is more stable numerically, but produces a Hessenberg matrix H, whose storage increases the memory required. However, if the order of H is kept moderately large, the storage memory may not be overwhelming. In order to retain good numerical stability properties, we propose a procedure based on the Arnoldi method with partial reorthogonalisation, combined with a Chebychev acceleration technique [2] [6].

2 The Arnoldi-Chebychev method with reorthogonalisation

2.1 The problem of the loss of orthogonality

Let us recall the incomplete Arnoldi algorithm. If A denotes a real or complex nonsymmetric matrix of order n, we seek a rectangular $n \times m$ ($m \leq n$) matrix V_m such that $V_m^* A V_m = H_m$ (where H_m is a Hessenberg matrix of order m) and $V_m^* V_m$ is the identity matrix of order n : the columns $v_1, \ldots, v_m$ of V_m form an orthonormal basis of the Krylov subspace $K_m = span\{v_1, Av_1, \ldots, A^{m-1}v_1\}$. This basis is built recursively by the following algorithm, using a given vector u. The projected Hessenberg matrix H_m is obtained at the same time .

Given $v_1 = \dfrac{u}{\|u\|_2}$,
  for $j = 1$ to m compute
    $h_j = V_j^* A v_j$, where $h_j^T = (h_{1k} \ldots h_{kk})$
    $x_j = Av_j - V_j V_j^* A v_j = (I_n - V_j V_j^*) A v_j$
    $h_{j+1,j} = \|x_j\|_2$
    $v_{j+1} = \dfrac{x_j}{h_{j+1,j}}$
  end

This is an implementation of the Gram-Schmidt algorithm at the step j, the vector Av_j is orthonormalised with respect to the columns of V_j. It is well-known that the Gram-Schmidt process can become unstable if these vectors are almost linearly dependent. This phenomenon, which also appears in other methods can implies incorrect eigenvalues approximations, by introducing "spurious" eigenvalues during the computation. We avoid this difficulty by using a recursive algorithm with partial reorthogonalisation [3] Since the dimension m of the Krylov subspace K_m is kept relatively small (< 1500), a standard procedure from LAPACK is used to compute the eigenvalues and the eigenvectors of H_m, denoted respectively by $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ and by $\tilde{y}_1, \ldots, \tilde{y}_m$. We then form the vectors $\tilde{x}_i = V_m \tilde{y}_i$, which are approximations to the eigenvectors of A associated with $\tilde{\lambda}_i$ (for $1 \leq i \leq m$). The normed residuals satisfy the relation

$$\|A\tilde{x}_i - \tilde{\lambda}_i \tilde{x}_i\| = h_{m+1,m} |e_m^T \tilde{y}_i|, \text{ where } e_m = (0, \ldots, 0, 1)^T.$$

Although they are mathematically equivalent, we will see in section 3 that the direct and the Arnoldi residuals respectively defined by

$$dr_i = \frac{\|A\tilde{x}_i - \tilde{\lambda}_i \tilde{x}_i\|}{\|\tilde{x}_i\|} \text{ and } ar_i = h_{m+1,m} \frac{|e_m^T \tilde{y}_i|}{\|\tilde{y}_i\|},$$

may not have the same numerical behaviour.

2.2 The iterative Arnoldi-Chebychev method

Let σ denote the set of the r wanted eigenvalues. Let τ denote the set of the remaining unwanted eigenvalues The spectrum of H_m is $sp(H) = \sigma \cup \tau$ The r desired eigenvalues can be either the dominant eigenvalues, or the r eigenvalues with greatest real or imaginary parts. The iterative Arnoldi Chebychev method is an hybrid algorithm, first proposed by Saad [6] The incomplete Arnoldi algorithm is restarted with a new vector of the form $z_k = p_k(A) z_0$ (where p_k is Chebychev a polynomial of degree $k-1$), in order to increase the rate of convergence of the r eigenvectors $\tilde{x}_i$ associated with the set σ. These vectors can be associated either with the r dominant eigenvalues, or with the r eigenvalues with greatest real or imaginary parts. The vector z_0 is taken as a linear combination of the r vectors $\tilde{x}_i$.
Because of the three terms recurrence satisfied by the Chebychev polynomials, the computation of the vector z_k is not expensive in terms of an arithmetic operations count.

3 Numerical results

We use the following notation .

- k is the degree of the Chebychev polynomial.

- m is the dimension of the Krylov subspace, i e. the order of the projected Hessenberg matrix H_m.
- n is the order of the matrix A.
- r is number of required ("wanted") eigenvalues (contained in the set σ).
- An Arnoldi-Chebychev step (abbreviated to "AC step") denotes the computation of a Hessenberg matrix H_m in the restarted Arnoldi algorithm.

The following example comes from a model of a plane in flight. The interesting modes of this system are described by complex eigenvalues whose imaginary parts lie in a frequency range chosen by the engineer. We demonstrate results for the computation of eigenvalues which have the largest imaginary parts. The matrix A is sparse with a block-structure. For large n, the matrix has many multiple and possibly defective eigenvalues. Figure 1 shows the general form of the spectra of these matrices — such spectra are typical in stability analysis. The Arnoldi-Chebychev

Aero example on Convex C-220 ($k = 200$ $m = 20$ $n = 2000$)				
Eigenvalue index	AC step	Residuals Arnoldi	Residuals direct	CPU time (s)
1	1	0.731 D-01	0.731 D-01	0.69
2		0.193 D+00	0.193 D+00	
1	2	0 305 D+01	0.305 D+01	0.67
2		0.467 D-03	0.467 D-03	
1	3	0 233 D-08	0,125 D-07	0.75
2		0 607 D-07	0 481 D-05	

Table 1: Aero example on Convex C-220 with parameters $k = 200$, $m = 20$, $n = 2000$ and $r = 4$

method is found to be very efficient when the order of the matrix is approximately less than 2000. Table 4 shows some results for a matrix of order 2000 with a ratio $\frac{m}{n} = \frac{1}{100}$ and $r = 4$, and where the stopping criterium is defined to be an Arnoldi residual less than or equal to 10^{-7}. In this case, we note that the direct and the Arnoldi residuals still agree. Unfortunately, when we increase the order of the matrix, we obtain unstable results — neither the Arnoldi residuals nor the direct residuals are correct. Note that the departure from normality increases rapidly with the order n, as shown on Figure 2. The eigenvalues are sorted by decreasing imaginary part and are represented by their index on the horizontal axis. The Arnoldi residuals that take the value zero (at machine precision) are not plotted. The non-zero Arnoldi residuals are plotted as functions of the eigenvalues index on the vertical axis. When $n = 8000, m = 800$ and $k = 200$, the direct residuals are also roughly constant (order 10^{12}), whilst the Arnoldi residuals have a very oscillatory behaviour as shown on Figure 3.

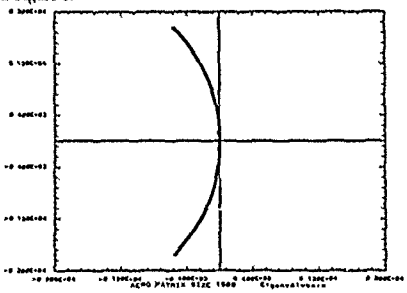


Figure 1: The spectrum of the matrix Aero for $n = 1500$

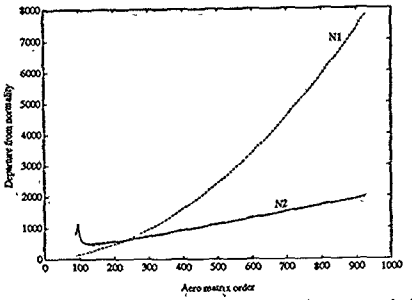


Figure 2: Departure from normality of the Aero matrix with the order of A variable from 90 to 925 with $N1 = \frac{\|AA^* - A^*A\|_F}{\|A\|_F}$ and $N2 = \frac{\|AA^* - A^*A\|_F}{\|A^2\|_F}$.

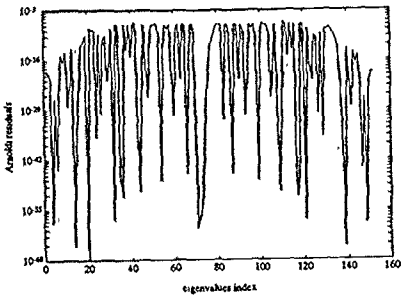


Figure 3: Arnoldi residuals for the first Arnoldi-Chebychev step with the parameters $k = 200$, $m = 800$, $n = 8000$, and $r = 4$

conclusion

The above experiments demonstrate that the departure from normality is an essential parameter when computing eigenvalues of nonsymmetric matrices. Further work is being carried out to understand the instabilities described above, notably the study of the departure of normality as a function of the order n

References

[1] F. Chatelin. *Valeurs propres de matrices* Masson, 1988

[2] F. Chatelin D. Ho and M. Bennani. Arnoldi-Chebychev method for large scale nonsymmetric matrices. *Math. Mod. and Num. Anal.*, 24:53-65, 1990.

[3] L. Kaufman J.W. Daniel, W.B. Gragg and G.W. Stewart. Reorthogonalization and Stable Algorithms for Updating the Gram-Schmidt QR Factorization. *Math. Comp.*, 30(136):772-95, 1976.

[4] Y. Saad. Chebychev Acceleration Techniques for Solving Nonsymmetric Eigenvalue Problems. *Math Comp*, 42(166) 567-88, 1984.

PHYSICAL SPACE REPRESENTATION OF SPECTRAL ENERGY TRANSFER IN HOMOGENEOUS TURBULENCE

J. ANDRZEJ DOMARADZKI
University of Southern California
Los Angeles, CA 90089-1191 U.S.A.

AND

ROBERT S. ROGALLO, ALAN A. WRAY
NASA-Ames Research Center
Moffett Field, CA 94035 U.S.A.

Abstract-Direct numerical simulations of homogeneous turbulence are used to analyze energy transfer among scales of motion in spectral space. A physical space representation of such a spectral energy transfer is devised and applied to the analysis of an eddy viscosity with a sharp spectral cut-off.

1 Introduction

Statistically homogeneous turbulent flows are conveniently represented in spectral (Fourier) space. In such a representation dynamically important, elementary nonlinear interactions involve three distinct modes with their wavenumbers forming a closed triad. Understanding these interactions is of paramount importance in the theory of turbulence since essentially all turbulence closures rely on assumptions about the nature of the nonlinear interactions. Recently, using results of direct numerical simulations (DNS) Domaradzki and Rogallo [1] [2] analyzed the energy transfer in homogeneous turbulence. They concluded that beyond the energy containing range the energy was transferred among scales of motion similar in size but that the interactions responsible for this local energy transfer were nonlocal in k-space. The importance of such nonlocal triadic interactions in the evolution of turbulent flows has been confirmed by Yeung and Brasseur [3] who also provided analytical arguments [4] supporting conclusions drawn from DNS.

Despite the usefullness of spectral representation as a theoretical and numerical tool in turbulence research, various quantities (velocity, energy, vorticity, etc.) in the physical space often provide a more natural description of turbulent flows. Thus it is of interest to have *the physical space* representation of the nonlinear transfer processes that dominate *the spectral space* dynamics. One such representation has been proposed by Domaradzki et al. [5]. In this paper we discuss other possible ways of representing detailed spectral energy transfer in the physical space.

2 Interscale Energy Transfer in Spectral Space

The equation for the energy amplitudes $\frac{1}{2}|u(\mathbf{k})|^2 = \frac{1}{2}u_n(\mathbf{k})u_n^*(\mathbf{k})$ is:

$$\frac{\partial}{\partial t}\frac{1}{2}|u(\mathbf{k})|^2 = -2\nu k^2\frac{1}{2}|u(\mathbf{k})|^2 + T(\mathbf{k}) \tag{1}$$

where $u_n(\mathbf{k})$ is the velocity field in spectral space, with the explicit dependence on time omitted, the asterisk denotes complex conjugate, ν is the kinematic viscosity, and $T(\mathbf{k})$ is the nonlinear energy transfer

$$T(\mathbf{k}) = Re(u_n^*(\mathbf{k})N_n(\mathbf{k})). \tag{2}$$

In the last equation $N_n(\mathbf{k})$ is the nonlinear term in the Navier-Stokes equation

$$N_n(\mathbf{k}) = (-i/2)P_{nlm}(\mathbf{k})\int d^3p u_l(\mathbf{p})u_m(\mathbf{k}-\mathbf{p}), \tag{3}$$

where tensor $P_{nlm}(\mathbf{k})$ accounts for the pressure and incompressibility effects. The summation convention is assumed throughout.

Detailed energy transfer to/from mode $\mathbf{k}$ caused by its interactions with wavenumbers $\mathbf{p}$ in a prescribed region $\mathcal{P}$ of the wavenumber space and $\mathbf{q} = \mathbf{k} - \mathbf{p}$ in another region $\mathcal{Q}$ is

$$T^{\mathcal{PQ}}(\mathbf{k}) = Re(u_n^*(\mathbf{k})N_n^{\mathcal{PQ}}(\mathbf{k})) \tag{4}$$

where $N_n^{\mathcal{PQ}}$ is (3) calculated with one of the contributing velocity fields truncated to $\mathcal{P}$ and the other to $\mathcal{Q}$. Details of such calculations are provided in [2]. For homogeneous turbulence the regions $\mathcal{P}$ and $\mathcal{Q}$ are usually chosen as spherical wavenumber bands. Similarly truncating velocity $u_n^*(\mathbf{k})$ in (4) to a spherical shell $\mathcal{K}$ results in a quantity $T^{\mathcal{KPQ}}(\mathbf{k})$ which, after averaging over $\mathcal{K}$, is interpreted as the energy transfer to the band $\mathcal{K}$ resulting from nonlinear interactions of scales in $\mathcal{K}$ with scales in $\mathcal{P}$ and $\mathcal{Q}$.

3 Interscale Energy Transfer in Physical Space

Inverse Fourier transform, signified by tilde, of $N_n(\mathbf{k})$ is the sum of the convection and pressure terms in the Navier-Stokes equation in the physical space coordinates

$$\tilde{N}_n(\mathbf{x}) = -\tilde{u}_i(\mathbf{x})\frac{\partial \tilde{u}_n(\mathbf{x})}{\partial x_i} - \frac{\partial p(\mathbf{x})}{\partial x_n}. \tag{5}$$

Similarly, using $N_n^{\mathcal{PQ}}(\mathbf{k})$ we can define its physical space counterpart $\tilde{N}_n^{\mathcal{PQ}}(\mathbf{x})$ as well as $\tilde{N}_n^{\mathcal{KPQ}}(\mathbf{x})$ which is the inverse Fourier transform of $N_n^{\mathcal{PQ}}(\mathbf{k})$ truncated to the band $\mathcal{K}$. $\tilde{N}_n^{\mathcal{PQ}}(\mathbf{x})$ can be interpreted as the contribution to the rate of change of the velocity field $\tilde{u}_n(\mathbf{x})$ at a point $\mathbf{x}$ made by the nonlinear interactions involving modes from the bands $\mathcal{P}$ and $\mathcal{Q}$. Note that these interactions influence all modes $\mathbf{k}$ which can form a triangle with two other modes such that one is in $\mathcal{P}$ and the other in $\mathcal{Q}$. $\tilde{N}_n^{\mathcal{KPQ}}(\mathbf{x})$ represents a contribution to the rate of change of $\tilde{u}_n(\mathbf{x})$ which is made by all modes from $\mathcal{K}$ interacting nonlinearly with modes in $\mathcal{P}$ and $\mathcal{Q}$.

The rate of change of the turbulent energy $e(\mathbf{x}) = \frac{1}{2}\tilde{u}_n(\mathbf{x})\tilde{u}_n(\mathbf{x})$ at a point $\mathbf{x}$ caused by the nonlinear interactions is

$$\frac{\partial e(\mathbf{x})}{\partial t} = \tilde{u}_n(\mathbf{x})\tilde{N}_n(\mathbf{x}). \tag{6}$$

Our goal is to decompose (6) into contributions from the interactions among modes from predefined wavenumber bands $\mathcal{K}, \mathcal{P}$, and $\mathcal{Q}$ i.e. to find a physical space counterpart of $T^{\mathcal{KPQ}}(\mathbf{k})$ which itself is the result of such a decomposition of the transfer $T(\mathbf{k})$ performed in the spectral space. Despite uniqueness of such a decomposition in the spectral representation, the procedure is ambiguous in the physical space. Possible definitions are:

$$\tilde{T}_1^{\mathcal{KPQ}}(\mathbf{x}) = \tilde{u}_n^{\mathcal{K}}(\mathbf{x})\tilde{N}_n^{\mathcal{PQ}}(\mathbf{x}), \tag{7}$$

$$\tilde{T}_2^{\mathcal{KPQ}}(\mathbf{x}) = \tilde{u}_n^{\mathcal{K}}(\mathbf{x})\tilde{N}_n^{\mathcal{KPQ}}(\mathbf{x}), \tag{8}$$

$$\tilde{T}_3^{\mathcal{KPQ}}(\mathbf{x}) = \tilde{u}_n(\mathbf{x})\tilde{N}_n^{\mathcal{KPQ}}(\mathbf{x}), \tag{9}$$

where $\tilde{u}_n^{\mathcal{K}}(\mathbf{x})$ is the inverse Fourier transform of $u_n(\mathbf{k})$ truncated to the band $\mathcal{K}$.

Function $\tilde{T}_1^{\mathcal{KPQ}}(\mathbf{x})$ is a straightforward counterpart of $T^{\mathcal{KPQ}}(\mathbf{k})$, with a product of $\tilde{u}_n^{\mathcal{K}}$ and $\tilde{N}_n^{\mathcal{PQ}}$ taken in the physical rather than in the spectral space. However, since $u_n^{\mathcal{K}}(\mathbf{k})$ vanishes outside $\mathcal{K}$, the multiplication in the spectral space implicitly truncates $N_n^{\mathcal{PQ}}(\mathbf{k})$ to the same band so that $T^{\mathcal{KPQ}}(\mathbf{k})$ expresses transfer to the modes in $\mathcal{K}$ only. In $\tilde{T}_1^{\mathcal{KPQ}}(\mathbf{x})$ the effect of nonlinear transfer to modes outside $\mathcal{K}$ is present in the term $\tilde{N}_n^{\mathcal{PQ}}$.

An explicit truncation of $N_n^{\mathcal{PQ}}(\mathbf{k})$ to $\mathcal{K}$ and multiplication by $\tilde{u}_n^{\mathcal{K}}$ seems to rectify this problem resulting in (8). The drawback of this definition is that it does not satisfy a natural condition:

$$\sum_{\mathcal{K},\mathcal{P},\mathcal{Q}} \tilde{T}^{\mathcal{KPQ}}(\mathbf{x}) = \frac{\partial e(\mathbf{x})}{\partial t}, \tag{10}$$

which is satisfied by both (7) and (9).

Function $\tilde{T}_3^{\mathcal{KPQ}}(\mathbf{x})$ may be interpreted as a fraction of the rate of change of the total energy $e(\mathbf{x})$ due to variation of modes in $\mathcal{K}$ as they are affected by nonlinear interactions with modes from $\mathcal{P}$ and $\mathcal{Q}$.

Thus none of the above definitions is an exact counterpart of the spectral transfer $T^{\mathcal{KPQ}}(\mathbf{k})$ but (9) is the most appealing candidate.

An interesting special case is obtained by dividing a wavenumber space into two disjoint regions $\mathcal{K}$ ($k \leq k_c$) and $\mathcal{P}$ ($k > k_c$). Quantity

$$T_{SGS}(\mathbf{x}|k_c) = \tilde{T}^{\mathcal{KPP}}(\mathbf{x}) + \tilde{T}^{\mathcal{K\bar{K}P}}(\mathbf{x}) \qquad (11)$$

provides a physical space representation of the rate of change of energy of large scales ($k \leq k_c$) due to their nonlinear interactions through wavenumber triads which have at least one of the legs in the region $\mathcal{P}$. This is precisely the energy transfer process which is the subject of the subgrid-scale modeling.

We have computed transfer functions (7) and (11) for the statistically isotropic velocity field obtained in direct numerical simulations performed with a resolution of 128^3 modes (maximum wavenumber $k = 64$). The low wavenumber band $\mathcal{Q}$ remains always fixed and is chosen to cover the entire energy containing range ($0 < q < 10$). Figure 1 shows one plane from the full transfer (7) representing in the physical space the energy transfer to eddies in the band $23 < k < 28$ caused by their interactions with eddies in the bands $20 < p < 25$ and $0 < q < 10$. The transfer function is spatially intermittent and is predominantly positive, indicating a flow of energy from the larger scales p to the smaller scales k.

We have attempted to correlate this physical energy transfer with a number of simpler quantities (rate-of-strain, dissipation, energy, etc.) calculated from the velocity field truncated in such a way as to contain only either large or small scales. We found that the energy of the velocity field truncated to large scales $0 < k < 10$ correlates very well with the energy transfer among small scales shown in figure 1. Correlation of other calculated quantities with the energy transfer, notably the square of the rate-of-strain tensor, was generally much worse. Therefore we conclude that the energy transfer among small scales occurs mostly at those physical locations which contain large amounts of turbulent energy rather than at the locations of high strain rate, an unexpected result. Indeed, until this paradox is resolved, we can not be confident that the particular measure of energy transfer that we have used is the appropriate one.

We have used formula (11) to calculate subgrid-scale (SGS) energy transfer for the same field with the cutoff wavenumber $k_c = 10$. A plane from the full SGS transfer field is plotted in figure 2. The transfer is characterized by the presence of both negative and positive regions. These indicate energy flux from and to the large scales respectively due to their interactions with the smaller scales. Standard subgrid-scale eddy viscosity models predict transfer in one direction only, from large to small scales.

4 Conclusions

We have devised a physical space representation of the energy transfer processes among scales of motion belonging to three distinct wavenumber bands in the spectral space and conclude from it that the energy transfer among small scales is highly intermittent in the physical space and correlates well with regions of significant large-scale energy.

As a particular case we have calculated a subgrid scale energy transfer in isotropic turbulence. The SGS transfer exhibits regions of energy drain from large to small scales as well as significant regions of reversed energy transfer from small to large scales. Classical eddy viscosity models assume that transfer is always from large to small scales, contrary to the results of direct calculations.

Acknowledgments-Work of one of the authors (JAD) was supported by the AFOSR Contract No. 90-0300 and by the NASA-Ames/Stanford Center for Turbulence Research.

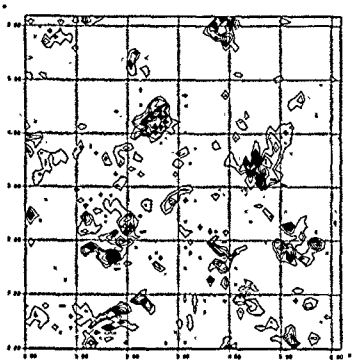


Fig.1. Energy transfer in physical space $\tilde{T}^{\mathcal{KPQ}}$ for $23 < k < 28, 20 < p < 25, 0 < q < 10$.

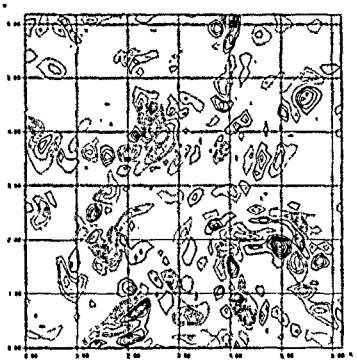


Fig 2. Subgrid scale energy transfer in physical space $T_{SGS}(\mathbf{x}|k_c)$ for $k_c = 10$.

References

[1] J.A. Domaradzki and R.S. Rogallo, in *Proceedings of Center for Turbulence Research, Summer Program 1988.*

[2] J A. Domaradzki and R.S. Rogallo, *Phys. Fluids* A2, 413 (1990).

[3] P.K. Yeung and J.G. Brasseur, submitted to *Phys. Fluids* A.

[4] J.G. Brasseur and P.K. Yeung, *AIAA Paper No. 91-0230* (1991)

[5] J.A. Domaradzki, R S. Rogallo and A A Wray, in *Proceedings of Center for Turbulence Research, Summer Program 1990.*

A METHOD TO EFFICIENTLY SOLVE TWO-POINT CLOSURE EQUATIONS FOR ANISOTROPIC TURBULENCE

YUKIO KANEDA
Department of Applied Physics, Nagoya University,
Chikusa-ku, Nagoya 464-01, Japan

Abstract A discussion is made of a method which may economize the computation to solve two-point closure equations for anisotropic turbulence. A method based on generating random fields to evaluate a certain wavevector integral and a time integral is presented. The use of Gaussian fields for the former has been so far found to yield a satisfactory approximation when applied to a closure equation for two dimensional Navier-Stokes(NS) turbulence.

I. INTRODUCTION

Among attempts to construct a statistical theory of turbulence in terms of low order moments, there are approaches based on renormalized perturbative expansions. Some of the so-called two-point or spectral closures thus obtained have been found to be in good agreement with experiments.

Although these closures are in principle applicable to a wide class of turbulences, there are algebraic and computational difficulties in solving closure equations applied to realistic anisotropic turbulences. One way to make closures more accessible to a wide class of turbulences is to develop an efficient method to solve such closure equations.

We present here a method based on the evaluation of integrals by generating random fields. We apply this method to the Lagrangian renormalized approximation (LRA)[1]. We confine ourselves here to the equation for single-time moments in the LRA. The structure of this equation is essentially similar to those in various closures including the direct-interaction approximation (DIA)[2], the abridged - Lagrangian - history - DIA[3], and the local-energy-transfer theory[4], and the method discussed here is in principle applicable not only to the LRA but also to these closures.

II. EQUATION FOR SINGLE-TIME MOMENTS

The LRA is a two-point closure approximation obtained by a simple truncation of Lagrangian renormalized perturbative expansions, and is free from any ad-hoc adjusting parameter. The LRA has been so far found in good agreement with experiments and direct numerical simulations regarding not only single-time quantities but also the two-time Lagrangian velocity auto-correlation function [5,6,7].

Let us consider homogeneous turbulence of an incompressible fluid of unit density obeying the NS equations with zero mean flow and no external force. Let $\bar{Q}$ be the Lagrangian covariance defined by

$$\bar{Q}_{ij}(\mathbf{x},t,\mathbf{x}',t') \equiv < \hat{v}_i(\mathbf{x},t'|t)\hat{v}_j(\mathbf{x}',t'|t') >, \quad (t \geq t'),$$

in which $\hat{\mathbf{v}}(\mathbf{x},t|s)$ is the velocity at time s of the fluid element that was at $\mathbf{x}$ at time t, and let $Q_{ik}(\mathbf{k}) = P_{ij}(\mathbf{k})\,\bar{Q}_{jk}(\mathbf{k})$ where $P_{ij}(\mathbf{k}) = \delta_{ij} - k_i k_j/k^2$,

$$\bar{Q}_{ij}(\mathbf{k},t,t') = (2\pi)^{-D}\int d^D\mathbf{r}\bar{Q}_{ij}(\mathbf{x}+\mathbf{r},t,\mathbf{x},t')e^{-i\mathbf{k}\cdot\mathbf{r}},$$

D is the dimension of space, and we use the summation convention for the repeated indices.

In the LRA, the evolution of the single-time moment $Q_{ij}(\mathbf{k};t;t)$ obeys

$$\left(\frac{\partial}{\partial t} + 2\nu k^2\right)Q_{ij}(\mathbf{k},t,t) = D_{ij}(\mathbf{k},t) + D_{ji}(-\mathbf{k},t),$$

$$D_{ij}(\mathbf{k},t) = \int_{t_0}^{t} ds T_{ij}(\mathbf{k},t,s), \tag{1}$$

$$T_{ij}(\mathbf{k},t,s) = \sum_{\mathbf{p},\mathbf{q}}^{\Delta} M_{imn}(\mathbf{k}) \times \{4M_{bca}(\mathbf{p})G_{mb}(\mathbf{p},t,s)Q_{nc}(\mathbf{q},t,s)Q_{ja}(-\mathbf{k},t,s) - 2M_{abc}(\mathbf{k})G_{ja}(-\mathbf{k},t,s)Q_{mb}(\mathbf{p},t,s)Q_{nc}(\mathbf{q},t,s)\},$$

$$M_{imn}(\mathbf{k}) = -(i/2)[k_m P_{in}(\mathbf{k}) + k_n P_{im}(\mathbf{k})], \tag{2}$$

where ν is the fluid viscosity, t_0 is the initial time, the symbol $\sum_{\mathbf{p},\mathbf{q}}^{\Delta}$ denotes the integrals over $\mathbf{p}$ and $\mathbf{q}$ satisfying $\mathbf{k} = \mathbf{p} + \mathbf{q}$,

$$Q_{ij}(\mathbf{k},t,s) = G_{im}(\mathbf{k},t,s)Q_{mj}(\mathbf{k},s,s), \tag{3}$$

and G is the so-called Lagrangian response function. We omit here writing the LRA equation for G, for which the reader may refer to Ref.1.

III. MONTE CARLO METHOD

Let us consider the evaluation of T_{ij} defined by (2), for given Q and G. Let ξ be a Gaussian random field with zero mean and covariance

$$< \xi_i(\mathbf{k},s)\xi_j(\mathbf{p},s) >= \delta_{\mathbf{k}+\mathbf{p}}Q_{ij}(\mathbf{k},s,s),$$

and let $\xi_i(\mathbf{k},t) = G_{ij}(\mathbf{k},t,s)\xi_j(\mathbf{k},s)$, then we have, because of (3),

$$< \xi_i(\mathbf{k},t)\xi_j(\mathbf{p},s) >= \delta_{\mathbf{k}+\mathbf{p}}Q_{ij}(\mathbf{k},t,s), \quad (t \geq s). \tag{4}$$

If we define z as

$$z_i(\mathbf{k}) = G_{im}(\mathbf{k},t,s)\sum_{\mathbf{p},\mathbf{q}}^{\Delta} M_{mjk}(\mathbf{k})\,\xi_j(\mathbf{p},s)\xi_k(\mathbf{q},s),$$

then

$$T_{ij}(\mathrm{k},t,s) = <\{\sum_{\mathrm{p,q}}^{\Delta} M_{imn}(\mathrm{k})\xi_m(\mathrm{p},t)\xi_n(\mathrm{q},t)\}z_j(-\mathrm{k})$$

$$+2<\{\sum_{\mathrm{p,q}}^{\Delta} M_{imn}(\mathrm{k})z_m(\mathrm{p})\xi_n(\mathrm{q},t)\}\xi_j(-\mathrm{k},t)> . \quad (5)$$

Now we introduce an approximation for the average $<f>$ of a random field f that

$$<f> \approx <f>_M, \quad (6a)$$

where

$$<f>_M \equiv \frac{1}{M_{max}}\sum_{m=1}^{M_{max}} f_m, \quad (6b)$$

f_m represents the value of f by the m-th realization of the Gaussian random field ξ, and M_{max} is the total number of the realizations, i.e., we approximate the average $<>$ in (5) by the Monte Carlo average $<>_M$ of finite number of realizations.

An approximation of (1) may be obtained by

$$D_{ij}(\mathrm{k},t) \approx \sum_{n=0}^{N} T_{ij}(\mathrm{k},t,s_n)\Delta s_n,$$

and by evaluating T_{ij} with (5) and (6) at each s_n, where $s_0 = t_0$, $s_N = t$, and Δs_n is an appropriate weight corresponding the discretization of the time interval.

The above Monte Carlo method has been tested by applying it to the LRA equations for two dimensional isotropic and anisotropic NS turbulences [8]. In practice the computation was simplified by using the stream function and vorticity formulation. It has been found that the values of the energy spectrum and the Lagrangian velocity auto-correlation obtained by using this method are in good agreement with the LRA solution and direct numerical simulations, even with $M_{max} = 1$.

The value of the integral in (2) over p, q for given values of the subscripts depends not only on i,m,n but also on the other subscripts a,b,c,j, so that the use of (2) in general requires us to evaluate the integrals for all possible values of the subscripts. On the other hand the integrals in (5) depend only on i,m,n and not on the other extra subscripts. Thus the evaluation in (5) is in general much easier than in (2), in particular . . three dimensions.

The above method is applicable, at least in principle, not only to NS turbulence with zero mean flow, but also to turbulent shear flow, Rossby turbulence, MHD turbulence, etc.. In its application to inhomogeneous turbulence, we need generate random Gaussian fields satisfying equations like

$$<\xi_i(\mathrm{k},t)\xi_j(\mathrm{p},s)> = Q_{ij}(\mathrm{k},\mathrm{p},t,s), \quad (7)$$

for given $Q_{ij}(\mathrm{k};\mathrm{p};t,s)$, instead of (4). Under an appropriate discretization, this is equivalent to generating fields satisfying equations like $<\xi_m\xi_n> = Q_{mn}$ in a symbolic notation, where the subscripts m,n stand for the wavevector, time and i or j in (7) -h ξ-field may be obtained by putting $\xi_m = A_{mn}y_n$ here

$$A_{mj}A_{nj} = Q_{mn}, \quad (8)$$

and $< y_m y_n > = \delta_{mn}$. It is not difficult to find A satisfying (8) by imposing $A_{mn} = 0$ for $m < n$.

The idea of using random fields may be applicable also to the time integration in (1). Suppose that the Gaussian random ξ-fields at time t and s are correlated as (4) for $s = s_n$, $n = 0,1,...,N$. Let $\{a\}$ and $\{b\}$ be sets of real random numbers independent of each other, and satisfying $<a_m a_n> = <(b_m - <b_m>)\, b_n> = \delta_{mn}$, $b_n > 0$ for any n,m, and let

$$\tilde{\xi}_i(\mathrm{k}) = \sum_{n=0}^{N} a_n\sqrt{b_n}\,\xi_i(\mathrm{k},s_n),$$

$$\tilde{G}_{ij}(\mathrm{k}) = \sum_{n=0}^{N}(b_n - <b_n>)G_{ij}(\mathrm{k};t,s_n)\Delta s_n,$$

$$\tilde{z}_i(\mathrm{k}) = \tilde{G}_{im}(\mathrm{k})\sum_{\mathrm{p,q}}^{\Delta} M_{mjk}(\mathrm{k})\,\tilde{\xi}_j(\mathrm{p})\tilde{\xi}_k(\mathrm{q}),$$

then the r.h.s. of (5) with z replaced by $\tilde{z}$ yields an approximation for $D_{ij}(\mathrm{k},t)$, (not T_{ij}). The average $<>$ may be then approximated by the Monte Carlo average $<>_M$ defined by (6) in which f_m depends not only on the realized random field ξ but also on $\{a\}$ and $\{b\}$. This method of evaluating D_{ij} has the advantage that it does not require the evaluation of integrals over p, q at each s_n. The performance/efficiency of this method remains to be tested.

ACKNOWLEDGMENT

I am indebted to the Research Foundation for the Electrotechnology of Chubu for a travel scholarship

REFERENCES

1. Y.Kaneda, J.Fluid Mech. 107,131(1981).
2. R.H.Kraichnan, J.Fluid Mech. 5,497(1959).
3. R.H.Kraichnan, Phys. Fluids 9,1728(1966).
4. W.D.McComb,J.Phys.A.Math.Gen.11,613(1978)
5. T.Gotoh,Y.Kaneda and N.Bekki, J.Phys.Soc.Jpn 57,866(1988).
6. Y.Kaneda and T.Gotoh, submitted to Phys. Fluids A.
7. T.Gotoh and Y.Kaneda, submitted to Phys. Fluids A.
8. Y.Kaneda, to be published in *Chaotic Dynamics and Transport in Fluids and Plasmas* (ed.V.Stefan, W.Horton,Y.Ichikawa,I.Prigogine & G.Zaslavsky). La Jolla International School of Physics Series: Research Trends in Physics, AIP(1991).

Stability, drag reduction and control of the turbulent boundary layer, using a low-dimensional model[1].

by

J. L. Lumley

Sibley School of Mechanical and Aerospace Engineering
Cornell University
Ithaca, NY 14853
USA

o0o

Abstract - Using an optimally convergent representation, a low dimensional model is constructed (Aubry *et al*, 1988) which embodies in a streamwise-invariant form the effects of streamwise structure (Berkooz *et al*, 1990). Results of Stone (1989) show that the model is capable of mimicking the stability change due to favorable and unfavorable pressure gradients. Results of Aubry *et al*, (1990) suggest that polymer drag reduction is associated with stabilization of the secondary instabilities, as has been speculated Results of Bloch & Marsden (1989) indicate that drag can be reduced by feedback, and that this is mathematically equivalent to polymer drag reduction.

Objective analysis of experimental measurements indicates that there are recurrent streamwise rolls present in the wall region of a turbulent boundary layer, at least in the quadratic mean sense (Corino & Brodkey, 1969, Kline *et al*, 1967). Representation theorems (Loève, 1955) permit optimal expansion of the instantaneous velocity field in the wall region in terms of these streamwise rolls (Lumley, 1967). Without involving ourselves in the question of the source of these rolls, we ask how they will behave dynamically. Severely truncating our system, and using Galerkin projection, we obtain a closed set of non-linear ordinary differential equations with ten degrees of freedom. The methods of dynamical systems theory are applied to these equations. Loss to unresolved modes is represented by a Heisenberg parameter (Aubry *et al*, 1988; Berkooz *et al*, 1990).

We find that for large values of the Heisenberg parameter (large loss), we obtain stable streamwise rolls having the experimentally observed spacing. For smaller values of the parameter, we have traveling waves (corresponding to cross-stream drift of the rolls), we also find a heteroclinic attracting orbit giving rise to intermittency; and finally a chaotic state showing ghosts of all of the above.

The intermittent jump in phase space from one attracting point to the other resembles in many respects the bursts observed in experiments. Specifically, the time between jumps, and the duration of the jumps, is approximately that observed in a burst, the jump begins with the formation of a narrowed and intensified updraft, like the ejection phase of a burst, and is followed by a gentle, diffuse downdraft, like the sweep phase of a burst. During the jump a spike of Reynolds stress is produced, as is observed in a burst, although the magnitude is limited in our model by the truncation of the high wavenumber components.

[1]Supported in part by the U. S. Air Force Office of Scientific Research under Contract No. AFOSR 89-0226, in part by the U. S. National Aeronautics and Space Administration, Langley Research Center, under Contract No., NAG-1 954, and in part by the U. S. National Science Foundation under Grants Nos. DMS-88-14553 and MSM 86-11164. Prepared for presentation at the 13th IMACS World Congress on Computation and Applied Mathematics, July 22-26, 1991, Trinity College, Dublin, Ireland.

The behavior is quite robust, much of it being due to the symmetries present (Aubry's group has examined dimensions up to 128 with persistence of the global behavior; Aubry & Sanghi, 1989). We have examined eigenvalues and coefficients obtained from experiment (Herzog, 1986), and from exact simulation (Moin, 1984), which differ in magnitude. Similar behavior is obtained in both cases; in the latter case, the heteroclinic orbits connect limit cycles instead of fixed points, corresponding to cross-stream waving of the streamwise rolls. The bifurcation diagram remains structurally similar, but somewhat distorted.

The role of the pressure term is made clear - it triggers the intermittent jumps, which otherwise would occur at longer and longer intervals, as the system trajectory is attracted closer and closer to the heteroclinic cycle. The pressure term results in the jumps occurring at essentially random times, and the magnitude of the signal determines the average timing. This clarifies the question of whether bursting scales with wall variables or with outer variables - evidently the structure of a burst scales with wall variables, while the time between bursts should scale in a complex way with both inner and outer variables (Stone & Holmes, 1990a, b).

Stretching of the wall region shows that the model is consistent with observations of polymer drag reduction (Aubry et al, 1990), in which one of the accepted mechanisms is the stabilization (by the extensional viscosity associated with the polymers) of the large eddies in the turbulent part of the flow, allowing the eddies to grow bigger and farther apart, as observed. Aubry et al (1989) tried stretching the eddy structure in the wall region, producing drag reduction, and found the bifurcation diagrams morphologically unchanged, except that the bifurcations occurred for larger and larger values of the Heisenberg parameter. This suggests that the motions giving rise to the bifurcations are more and more unstable, the more the region is stretched, requiring a larger and larger value of the Heisenberg parameter to stabilize them. Now, the Heisenberg parameter represents the loss of energy to the unresolved modes as well as loss to any other dissipation mechanism, such as viscosity or extensional viscosity. Hence the findings of Aubry et al (1989) are completely consistent with the idea of the larger eddies being less stable, and able to grow to this larger, less stable size due to the stabilizing effect of the polymer.

Change of the third order coefficients, corresponding to acceleration or deceleration of the mean flow, changes the heteroclinic cycles from attracting to repelling, increasing or decreasing the stability, in agreement with observations (Stone, 1989).

The existence of fixed points is an artifact introduced by the projection, in the exact equations, the rolls always decay after a period of growth. The projection incorporates statistically the birth and death of individual rolls, producing a stationary situation. However, a decoupled model still displays the rich dynamics (Moffat, 1989; Holmes, 1990a, Berkooz *et al*, 1990).

This sort of relatively simple model could be used as a "black box" in feed-back systems to control the boundary layer, as well as being used to predict pressure and stress fluctuations at the wall, and the effect of various drag reduction schemes. Feeding back eigenfunctions with the proper phase can delay the bursting, (the heteroclinic jump to the other fixed point), decreasing the drag. It is also possible to speed up the bursting, increasing mixing to control separation (Bloch & Marsden, 1989).

In recent work, Berkooz (1990), in collaboration with Holmes and Lumley, has shown that several assumptions made on an intuitive basis in the work of Aubry *et al* may be justified formally, namely: that the Heisenberg model used gives the correct dissipation within a constant of order unity, as assumed; that the Leonard stresses may be neglected in the case of modeling with no streamwise variation, as assumed; that the previous result holds for an arbitrary number of eigenfunctions when no streamwise variation is present; that models with no streamwise variation in effect average the streamwise dynamics, as conjectured by Holmes.

Early work of Bloch & Marsden (1989) showed that systems with homoclinic attractors are in principle controllable, with a certain type of control input. Our efforts in this direction are concentrated on determining the feasability of control, and trying to understand the possible gains in terms of drag reduction and mixing enhancement. To this end Berkooz, Holmes and Lumley introduced the notion of short term tracking time T_s. This is a measure of the time over which a dynamical systems model tracks the true dynamics accurately. T_s is of fundamental importance in the control application, and it must be of the order of the wall-region time scales to make control possible. They then showed that dynamical systems based on the Proper Orthogonal Decomposition have, on the average, the best T_s for a given number of modes (Berkooz, 1991).

Berkooz (1990) has made rigorous estimates using the proper orthogonal decomposition showing that a structured turbulent flow, such as the wall layer, has a phase space representation that remains within a thin slab centered on the most energetic modes for most of the time. However, exits from this region, which is all that our low-dimensional models include, should not be ignored, since they typically correspond to violent events, such as the bursting phenomenon. Berkooz and Holmes are trying to develop a theory in which deterministic, low-dimensional dynamics governing the low modes apply most of the time, passages from and returns to this being modeled probabilistically. This might be viewed as a dynamical closure. They plan to test their theory on problems including the 32 and 54 dimensional projections of Aubry & Sanghi (1989).

Campbell & Holmes (1990) are continuing their studies of symmetry breaking (O(2) → D4) in systems with structurally stable heteroclinic cycles. They have proved that no analytic (second) integral of motion exists in a certain limiting case and that only two pairs of the continuum of O(2) symmetric heteroclinic cycles persist in general. They are studying the bifurcations from these survivors. This work is relevant to our models of interacting coherent structures in boundary layers with discrete spanwise symmetry, such as that caused by riblets. This is to our knowledge the first analytical contribution to our understanding of the drag reduction caused by riblets.

BIBLIOGRAPHY.

Aubry, N. & Sanghi, S. 1989. Streamwise and cross-stream dynamics of the turbulent wall layer. Proceedings, July meeting of ASME, New York. ed Ghia.

Aubry, N., Holmes, P. & Lumley, J. L. 1990. The effect of modeled drag reduction on the wall region. *Theoretical and Computational Fluid Dynamics*, 1: 229-248.

Aubry, N., Holmes, P., Lumley, J. L. and Stone, E. 1988. The dynamics of coherent structures in the wall region of a turbulent boundary layer. *J. Fluid Mech.* 192: 115-173.

Berkooz, G. 1990. Observations on the proper orthogonal decomposition. In *The Lumley Symposium. Recent Developments in Turbulence*, eds. T. Gatski & C. Speziale. Berlin etc: Springer. To appear.

Berkooz, G. 1991. Private communication.

Berkooz, G. Gackenheimer, J., Holmes, P., Lumley, J.L., Marsden, J., Aubry, N. & Stone, E. 1990 Dynamical-systems theory approach to the wall region. AIAA Paper No. 90-1539.

Berkooz, G., Holmes, P. & Lumley, J. L. 1990. Control of the boundary layer and dynamical systems theory: an update. In *The Global Geometry of Turbulence*, ed. Javier Jiménez. Berlin etc: Springer. In press.

Berkooz, G., Holmes, P. & Lumley, J. L. 1990. Intermittent dynamics in simple models of the turbulent wall layer. *J. Fluid Mech.* Accepted for publication.

Berkooz, G., Holmes, P. and Lumley, J. L. 1989. Decaying turbulence and heteroclinic cycles in the wall layer.
Bloch, A. M. and Marsden, J. E. 1989. Controlling Homoclinic Orbits. *Theoretical and Computational Fluid Dynamics*. 1(3): 179-190.

Campbell, S. & Holmes, P. 1990. Bifurcation from O(2)-symmetric heteroclinic cycles with three interacting modes. *Nonlinearity*. To appear.
Corino, E. R., and Brodkey, R.S. 1969. A visual investigation of the wall region in turbulent flow. *J. Fluid Mech.* 37(1):1-30.
Herzog, S. 1986. *The large scale structure in the near-wall region of turbulent pipe flow*. Ph.D. thesis, Cornell University.

Holmes, P. J. 1990. Can dynamical systems approach turbulence? In *Whither Turbulence? Turbulence at the crossroads*. Lecture Notes in Physics Vol. 357 ed. J. L. Lumley, pp. 195-249 Berlin etc: Springer.

Holmes, P. J. 1990. Nonlinear dynamics, chaos and mechanics. *Applied Mechanics Reviews* 43 (5, 2): S23-S39.

Holmes, P. J. 1990. On Moffatt's paradox, or can empirical projections approach turbulence? Author's close-up. In *Whither Turbulence? Turbulence at the crossroads*. Lecture Notes in Physics Vol. 357 ed. J. L. Lumley. pp. 306-312 Berlin etc: Springer.

Kline, S.J., Reynolds, W.C., Schraub, F.A. and Rundstadler, P.W. 1967. The structure of turbulent boundary layers. *J. Fluid Mech.* 30(4): 741-773.
Loève, M. 1955. *Probability Theory*. New York: Van Nostrand

Lumley, J.L. 1967. The structure of inhomogeneous turbulent flows. In *Atmospheric Turbulence and Radio Wave Propagation*, A.M. Yaglom and V.I. Tatarski, eds.: 166-178. Moscow: Nauka.

Lumley, J.L. 1981. Coherent structures in turbulence. *Transition and Turbulence*, edited by R.E. Meyer, Academic Press, New York: 215-242.

Moffat, H. K. 1989 Fixed points of turbulent dynamical systems and suppression of non-linearity. In *Whither Turbulence*, ed. J. L. Lumley. Heidelberg: Springer. In press.

Moin, P. 1984. Probing turbulence via large eddy simulation. AIAA 22nd Aerospace Sciences Meeting.

Stone, E. & Holmes, P. 1990. Random perturbations of heteroclinic attractors. *SIAM J on Applied Math*. 50: 726-743.

Stone, E. & Holmes, P. 1990. Unstable fixed points, homoclinic orbits and exponential tails. *Phys. Lett. A* Submitted.

Stone, E. 1989. *A Study of Low Dimensional Models for the Wall Region of a Turbulent Boundary Layer*. Ph. D. Thesis. Ithaca, NY. Cornell University.

APPLICATION OF RENORMALISATION GROUP TO TURBULENCE SIMULATION USING CONDITIONAL AVERAGING.

W. D. McComb,
Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, U.K.

Turbulence presents the archetypal problem of the nonlinear field which exhibits chaotic behaviour. Above a critical value of the Reynolds number, the velocity field $U(x,t)$ varies unpredictably with position and time. On the other hand, if we Fourier transform the velocity with respect to wavenumber k, then the problem becomes one of many degrees of freedom: the Fourier modes $U(k,t)$, in the interval $0 \le k \le k_0$. The upper bound k_0 is defined through the dissipation integral

$$\epsilon = \int_0^\infty 2\nu_0 k^2 E(k)\,dk \simeq \int_0^{k_0} 2\nu_0 k^2 E(k)\,dk, \qquad (1)$$

where ϵ is the dissipation rate, ν_0 is the kinematic viscosity, and $E(k)$ is the energy spectrum. This definition ensures that k_0 is of the same order of magnitude as the Kolmogorov dissipation wavenumber.

In general, for all but the simplest turbulent fields and lowest Reynolds numbers, there are too many degrees of freedom for complete numerical simulation of the Navier-Stokes equation (NSE) to be possible. Computers are not large enough for this purpose, nor are they likely to be so in the foreseeable future. Thus the theorist is faced with the important and interesting question: can we devise some analytic method to reduce the number of modes which must be resolved by the computer? In this way we could hope to make calculation feasible in a hybrid fashion, with a solution to the problem which is neither wholly analytic nor wholly numerical.

The obstacle which lies in the way of such an approach is the well known phenomenon of nonlinear mixing. This may be seen as follows. Consider the solenoidal form of the NSE in wavenumber space. This takes the form [1]:

$$(\partial/\partial t + \nu_0)U_\alpha(\mathbf{k},t) = M_{\alpha\beta\gamma}(\mathbf{k}) \int d^3j\, U_\beta(\mathbf{j},t) U_\gamma(\mathbf{k}-\mathbf{j},t) + f_\alpha(\mathbf{k},t), \qquad (2)$$

where

$$M_{\alpha\beta\gamma}(\mathbf{k}) = (2i)^{-1}[k_\beta D_{\alpha\gamma}(\mathbf{k}) + k_\gamma D_{\alpha\beta}(\mathbf{k})], \qquad (3)$$

and the projection operator (which arises in the process of eliminating the pressure and the incompressibility condition together) is given by

$$D_{\alpha\beta} = \delta_{\alpha\beta} - k_\alpha k_\beta |\,\mathbf{k}\,|^{-2}, \qquad (4)$$

where $\delta_{\alpha\beta}$ is the Kronecker delta. We choose the stirring forces $f_\alpha(\mathbf{k},t)$ to satisfy the usual requirements for a well-posed problem. That is, they act directly on the fluid only at the lowest values of the wavenumber. Although they are therefore somewhat arbitrary in form, they are only introduced in order to allow us to consider the simplest situation, which is turbulence which is homogeneous, isotropic and stationary. We shall also restrict our attention to the case of zero mean field.

From equation (2) it is now obvious that the nonlinear term couples the modes corresponding to different wavenumbers together. In principle, it follows that all modes are coupled in this way, although in practice it is generally assumed that appreciable interaction only occurs between modes which are close to each other in wavenumber. Indeed, it is the assumption of locality of energy transfer in wavenumber which underpins the idea of an energy cascade and which leads, by way of dimensional analysis, to the experimentally observed Kolmogoroff spectrum; thus:

$$E(k) = \alpha \epsilon^{2/3} k^{-5/3}, \qquad (5)$$

where α is the constant of proportionality.

Now let us eliminate modes by the application of the Renormalisation Group. We begin by decomposing the velocity field into explicit scales for $k \le k_1$ and implicit scales for $k_1 \le k \le k_0$; thus:

$$\begin{aligned} U_\alpha(\mathbf{k}) &= U_\alpha^-(\mathbf{k}) \text{ for } 0 \le k \le k_1 \\ &= U_\alpha^+(\mathbf{k}) \text{ for } k_1 \le k \le k_0, \end{aligned} \qquad (6)$$

where k_1 is defined by

$$k_1 = (1-\lambda)k_0, \qquad (7)$$

with the bandwidth parameter λ satisfying the condition $0 \le \lambda \le 1$.

In principle, the renormalization-group approach can be carried out as follows:

(A) Solve the NSE on $k_1 \le k \le k_0$. Substitute that solution for the high-k modes back into the NSE on $0 \le k \le k_1$. This results in an increment to the viscosity: $\nu_0 \to \nu_1 = \nu_0 + \delta\nu_0$.

(B) Rescale the basic variables such that the NSE on $0 \le k \le k_1$ looks like the original Navier-Stokes equation on $0 \le k \le k_0$.

These two stages are then repeated to eliminate the effect of high wavenumbers progressively in a series of bands $k_{n+1} \le k \le k_n$, until the rescaled viscosity no longer changes (i.e. has reached a "fixed point").

This procedure is appealingly simple, and has a clear physical interpretation, but it has not proved easy to put it into practice in the turbulence problem. Certainly, the method of *iterative averaging* has shown that a fixed point may be obtained, with reasonable quantitative results [2, 3]. But, this method has been open to the criticisms that the basic averaging technique used was obscure and that there was an unexplained dependence on the bandwidth parameter λ. However, recently these criticisms have been answered by the introduction of a *conditional average*, which is then evaluated as an approximation in which λ plays the part of a small parameter [4]. It is this development which concerns us here.

Let us introduce a conditional average which smooths out the effect of the high-k modes, while keeping the U^- constant. We represent it by an operator $A[U^+ \,|\, U^-]$ and denote its effect on the first shell of wavenumbers to be eliminated by $<>_0$, thus:

$$A[U^+ \,|\, U^-]U_\alpha U_\beta ... U_\gamma = < U_\alpha U_\beta ... U_\gamma >_0 . \qquad (8)$$

It then follows that this operator has the properties:

$$< U_\alpha^-(\mathbf{k}) >_0 = U_\alpha^-(\mathbf{k}), \qquad (9)$$

$$< U_\alpha^-(\mathbf{j}) U_\beta^-(\mathbf{k}-\mathbf{j}) >_0 = U_\alpha^-(\mathbf{j}) U_\beta^-(\mathbf{k}-\mathbf{j}). \qquad (10)$$

It is at this point that we encounter the difficulties associated with nonlinear mixing. We now wish to evaluate averages of the above kind over the high-k modes and express them in terms of global mean quantities, such as the energy spectrum. Evidently the problem we face is that the U^+ field is not independent of

the U^- field which we are holding constant. The two fields are coupled together through the nonlinear term in equation (2).

We tackle this difficulty by writing the high-k modes in terms of a new field V_+; thus

$$U^+_\alpha(k,t) = V^+_\alpha(k,t) + \Delta^+_\alpha(k,t). \quad (11)$$

Here V^+ is a field of the same general type as U^+ and belongs to the same turbulent ensemble. Hence it has the same properties under *global* averaging; thus

$$< V^+_\alpha(k,t) > = 0, \quad (12)$$

$$< V^+_\alpha(k,t) V^+_\beta(k',t') > = < U^+_\alpha(k,t) U^+_\beta(k',t') > . \quad (13)$$

However, the essential feature of V^+ is that it is not coupled to the low-k modes. Thus, through (11) we introduce the function Δ^+ to take account of mode coupling. We note that it follows immediately from (12) that Δ^+ has zero mean under global averaging.

In order to complete our specification of the two new fields, we state their properties under *conditional* averaging as

$$A[U^+ \mid U^-]V^+(k,t) = 0 \quad (14)$$

and

$$A[U^+ \mid U^-]\Delta^+(k,t) = O(\lambda)^m, m \geq 1. \quad (15)$$

We now make use of the above equations to evaluate conditional moments of the U^+ in terms of unconditional moments of the V^+, and in this way we decompose the NSE into an equation for the explicit scales but containing the term $< U^+U^+ >_0$ as representing the coupling to implicit modes, along with an iterative solution for the latter quantity, which can be shown to be linearly dependent on $U^-(k,t)$. In this way, we can derive an expression of the increment to the fluid viscosity as the sum of an infinite series in the unconditional moments of the V^+ field.

In order to extend this calculation to progressively lower bands in wavenumber, we have to make three approximations. First, we evaluate convolution integrals over intermediate times on the assumption that the U^+ evolve much more rapidly than (on average) the U^-. Second, we truncate the expansion in moments of the U^+, on the grounds that the expansion is bounded by a power series in the velocity field in the dissipation range of wavenumbers. Thirdly, and this is the main new feature of the work, we relate the V^+ field to the U^+ by assuming that the coupling between distinct Fourier modes is local in wavenumber. That is, we assume that λ is large enough for $U(k_0)$ to be independent of $U(k_1)$, and at the same time that λ is small enough for us to represent the Fourier components in the band by means of a first-order Taylor series. In this way, we impose both upper and lower bounds on λ, when we make the identification

$$V^+_\alpha(k,t) = U^+_\alpha(k_0,t) + (k-k_0).\nabla U^+_\alpha(k,t)\mid_{k=k_0} + O(\lambda^2). \quad (16)$$

The RG procedure may then be implemented band by band, with the scaling transformation

$$k_{n+1} = (1-\lambda)k_n = hk_n, \quad k = k_{n+1}k', \quad (17)$$

for each iteration n. An assumption that the energy spectrum is a power law, followed by power counting, leads to the following expression for the effective viscosity

$$\nu(k_n k') = \alpha^{1/2}\epsilon^{1/3}k_n^{-4/3}\tilde{\nu}_n(k'), \quad (18)$$

where the coefficient $\tilde{\nu}_n(k')$ is given by the recursion relationship

$$\tilde{\nu}_{n+1}(k') = h^{4/3}\tilde{\nu}_n(hk') + h^{-4/3}\delta\tilde{\nu}_n(k'), \quad (19)$$

with

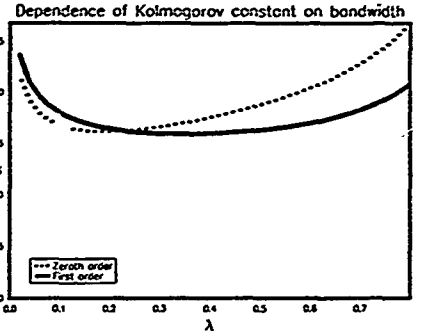
$$\delta\tilde{\nu}_n(k') = \frac{1}{4\pi k'^2}\int_{\mathcal{S}} d^2\vec{j}\,\frac{I(k',\vec{j})Q'}{\tilde{\nu}_n(h\vec{j})j^{2/3} + \tilde{\nu}(k')k'^{2/3}}, \quad (20)$$

for the wavenumber bands $0 \leq k' \leq 1, 1 \leq \vec{j}, \vec{r} \leq h^{-1}$, where $\vec{r} = |\vec{k}' - \vec{j}'|$, and

$$Q' = k'^{11/3} - \frac{11}{3}k'^{14/3}(\vec{r} - k^{-1}) + \ldots, \quad (21)$$

The coefficient $I(k',\vec{r})$ contains purely geometrical factors and is given in detail in [4].

As a test of the method, we have calculated the Kolmogorov spectral constant, as defined by equation (5). Over a range of values of λ for which the theory is valid, we find $\alpha = 1.6$, in good agreement with experiment. At large values of λ, one may observe the breakdown of the first-order Taylor series approximation; whereas, at small values one sees the effects of mode coupling, which would invalidate the basic assumption that $U(k_0)$ is independent of $U(k_1)$.



References

[1] W. D. McComb, *The Physics of Fluid Turbulence*, Clarendon Press, Oxford, (1990).

[2] W. D. McComb, Phys. Rev. A 26, 1078 (1982).

[3] W. D. McComb, In *Direct and large-eddy simulation* (eds. U. Schumann and R. Friedrich), *Notes on numerical fluid mechanics*, Vol 15, Vieweg, Braunschweig (1986).

[4] W. D. McComb and A. G. Watt, Phys. Rev. Lett. 65, 3281 (1990).

Phase Correlations and Transfer of Energy and Helicity.[1]

Yoshiki MURAKAMI
Department of Mathematical Sciences,
College of Engineering,
University of Osaka Prefecture, Sakai, 591, Japan.

Abstract Three kinds of random transformation which preserves the amplitude of each mode, $|\vec{v}(\vec{k})|^2$ are applied to several ranges in Fourier space of turbulent fields which are obtained numerically at relatively low Reynolds number ($R_\lambda \approx 30$). The effects of destroying the correlations between modes using these methods to both energy and helicity transfer are investigated. It is shown that one method which also preserves the helicity spectrum is the most effective to change the form of both transfer spectra and that it works as if it were mode-filtering.

I. INTRODUCTION

We investigate the role of phase correlations on the nonlinear transfer in the universal range. We treat the following problem: *If we destroy the correlations between modes in some scales keeping the energy spectral form artificially, how are the efficiency of the energy transfer (i.e. energy cascade) effected?* This destruction may be regarded as an idealization of the external random disturbances which inevitably exist in laboratory experiments and natural phenomena. Hence this problem is related to the *structural* stability problem on the turbulent equilibrium state. Along this line, we have investigated various aspects of the *dynamics* of the decaying disturbed turbulence[1,2]. Here we do not consider the dynamics but apply reshufflings to various ranges of Fourier space in order to estimate the effect to not only the energy transfer but also the helicity transfer which was not treated in Murakami *et al*[2].

II. NUMERICAL PROCEDURE

We used the data obtained by direct numerical simulations of the incompressible Navier-Stokes (NS) equation with periodic boundary conditions by pseudospectral method ($N = 64^3, \nu = 0.015$). Details of the numerical scheme are given in Polifke and Shtilman[3] and Polifke[4]. One field (RH) corresponds to the statistically steady state generated by a time-independent and *random* helicity forcing exerted at large scales ($3 < k < 4$). The field has been obtained after around six turnover time (1000 time steps with $\Delta t = 0.001$). The other(MH) is the same as the above except using *maximum* helicity forcing. The characteristic parameters of these fields are given in the Table I. Note that we treat the relatively low Reynolds number ($R_\lambda = 26.5$ and 33.4) turbulence owing to the limitation of the resolution.

TABLE I The characteristic parameters of the turbulent fields. E: total energy, Ω: total enstrophy, H: total helicity, $R_\lambda = \sqrt{10/3}E/(\nu\sqrt{\Omega})$: Taylor's micro-scale Reynolds number, $\tau = (\frac{3}{4}\pi \int k^{-1}E(k)dk)\,/\sqrt{2E/3}$: turnover time.

TABLE I

Field	E	Ω	H	R_λ	τ
(a) RH	1.29	35.0	0.79	26.5	0.70
(b) MH	1.43	27.8	8.47	33.4	0.71

[1]This work was performed with partial support of U S Department of Energy Grant No DE-FG02ER13837 All computations were performed at Lawrence Livermore National Laboratory

We give the definition of three different methods for *energy-conserving* transformation, which we call reshuffling. ($E'(\vec{k}) = E(\vec{k})$):[2]

(1) *HR method.* Helicity spectrum is given by $H(\vec{k}) = \vec{v}(\vec{k})\cdot\vec{\omega}(-\vec{k}) = 2|\vec{k}||\mathrm{Re}\vec{v}(\vec{k})||\mathrm{Im}\vec{v}(\vec{k})|\sin\phi(\vec{k})$, where $\phi(\vec{k})$ is the angle between these two vectors (the helicity associated phase). We only change the orientation of the real part, $\theta(\mathrm{Re}\vec{v}(\vec{k}))$ in a uniformly random way in all directions $[0-2\pi]$. $[|\mathrm{Im}\vec{v}'(\vec{k})| = |\mathrm{Im}\vec{v}(\vec{k})|$ and $|\mathrm{Re}\vec{v}'(\vec{k})| = |\mathrm{Re}\vec{v}(\vec{k})|]$

(2) *HC method.* We preserve the angle, $\phi(\vec{k})$ between real and imaginary part of $\vec{v}(\vec{k})$, but rotate this pair in a uniformly random way, thus preserving both energy $E(\vec{k})$ and helicity $H(\vec{k})$. $[|\mathrm{Re}\vec{v}'(\vec{k})| = |\mathrm{Re}\vec{v}(\vec{k})|$, $|\mathrm{Im}\vec{v}'(\vec{k})| = |\mathrm{Im}\vec{v}(\vec{k})|$ and $\phi'(\vec{k}) = \phi(\vec{k})$.]

(3) *SA method.* We apply the following transformation:

$\mathrm{Re}\vec{v}'(\vec{k}) = (\sqrt{\alpha E(\vec{k})}/|\mathrm{Re}\vec{v}(\vec{k})|)\mathrm{Re}\vec{v}(\vec{k})$ and $\mathrm{Im}\vec{v}'(\vec{k})$

$= (\sqrt{(1-\alpha)E(\vec{k})}/|\mathrm{Im}\vec{v}(\vec{k})|)\mathrm{Im}\vec{v}(\vec{k})$, where α is a uniformly random number between 0 and 1. $[\theta(\mathrm{Re}\vec{v}'(\vec{k})) = \theta(\mathrm{Re}\vec{v}(\vec{k}))$, $\theta(\mathrm{Im}\vec{v}'(\vec{k})) = \theta(\mathrm{Im}\vec{v}(\vec{k}))$ and $E'(\vec{k}) = E(\vec{k})$.]

We apply each reshuffling separately and calculate both energy and helicity transfer and compare the results on the undisturbed case and the filtered spectrum where some modes in a range are set to zero. We report the results on two ranges. $15 < k < 32$ and $6 < k < 11$. to two fields: (a) random helicity field (RH) and (b) maximum helicity field (MH).

III. RESULTS

A. Turbulent fields

In Fig. 1 both energy and helicity spectra for (b) MH is shown. The dotted lines shows $\sigma_G(H(\vec{k})) = (1/\sqrt{\pi})E(k)$ which is the standard deviation of the helicity spectrum in the quasi-Gaussian approximation[3,4] while the solid line shows the maximum helicity $2kE(k)$. Positive values of the helicity are denoted by plus signs, negative by diamonds. There is no clear inertial range owing to the relatively low Reynolds number. (See TABLE I.)

In case (a) RH almost all $H(k)$ lie within quasi-Gaussian, whose figure is omitted, while clear deviation is observed in case (b) MH.

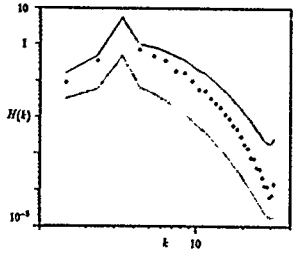


FIG. 1. Helicity cospectrum and energy spectrum in log-log scale for case (b) MH.

We recall both balance equations for the energy and helicity spectra

$$\frac{dE(k)}{dt} = T_E(k) - 2\nu k^2 E(k) \qquad (1)$$

and

$$\frac{dH(k)}{dt} = T_H(k) - 2\nu k^2 H(k), \qquad (2)$$

where we omit the term due to the forcing ($3 < k < 4$). The transfer terms $T_E(k)$ and $T_H(k)$ are defined as the shell-averaged of

$$T_E(\vec{k}) = Re[\vec{\lambda}(\vec{k}) \cdot \vec{v}^*(\vec{k})] \tag{3}$$

and

$$T_H(\vec{k}) = 2\vec{k} \cdot Im[\vec{\lambda}^*(\vec{k}) \times \vec{v}(\vec{k})]. \tag{4}$$

where $\vec{\lambda}(\vec{k})$ is a Fourier transform of the Lamb vector, $\vec{\lambda}(\vec{x}) = (\vec{v} \times \vec{\omega})(\vec{x})$ and the asterisk denotes the complex conjegate. Note that the reshuffilngs keep $E(\vec{k})$ but destroy $T_E(\vec{k})$ and $T_H(\vec{k})$ and the integrals of $T_E(\vec{k})$ and $T_H(\vec{k})$ over all modes are also zero after transformation. Eq. (4) tells that $T_H(\vec{k})$ is also defined as the projection of $\vec{\lambda}(\vec{k})$ in this plane as well as $T_E(\vec{k})$. Hence the nonlinear reduction mechanism observed numerically by Shtilman and Polifke[5], the tendency of the alligament of $\vec{\lambda}(\vec{k})$ and $\vec{k}$, is also essential to the *helicity* transfer.

We touch on the form of the transfer spectra in the dissipation range. In the steady state we obtain the following inequality:

$$|T_H(\vec{k})| \leq 2kT_E(\vec{k}), \tag{5}$$

where the equality holds for maximum helicity modes irrespective of the steady condition. Note that the inequality does not hold generally. In our fields eq. (5) is satisfied with both cases ($[T_H(k)]/(2kT_E(k))$ is the order of .01 for (a) and 0.1 for (b)) although the steady condition is not satisfied.

B. Energy and helicity transfer spectra

All cases in Fig. 2 where we treat the energy transfer for case (a) RH show that the HC method is the most effective of three and the curve S is very close to the curve N and that the HC method is almost the same as the filtered case (curve F). In Fig. 2 ($15 < k < 32$) the reshuffled cases (the curves R,C and S) and the filtered case (the curve $\bar{F}$) have a peak (the energy barrier) at the boundary between the reshuffled region and the unreshuffled one. The reshuffled region ($k > 15$) of the HC is destroyed completely as that of the filtered case. Note that the definition of $\vec{\lambda}(\vec{k})$ depends on the correlations between *all* scales. On the other hand, five curves are closer to each other in the larger scales ($k < 10$). The small scale reshuffling does not affect the larger scales noticeably. In the case ($6 < k < 11$), whose figure is omitted, in the small scales next to the reshuffled region of the curves C and F we observe the hollows in contrast the curves N and S. The HC works as if it cut the larger scales, $0 < k < 5$. We stress that the results for case (b) MH is almost the same as the above; the helicity does not affect basic mechanism of the correlations which we treat here.

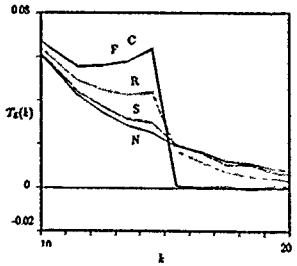


FIG. 2. Energy transfer spectra in lin-lin scale for (a) RH. The symbol N indicates the nondisturbed case; R: HR, C: HC, S: SA and F. Filtered case. These notations are used throughout this paper. N: Solid line, R: Dotted line, C: Dashed line, S: Chain-dotted line and F: Solid line. $15 < k < 32$.

Fig. 3 shows the helicity transfer spectra for (a) RH ($15 < k < 32$). We cannot see a clear peak at the boundary between the reshuffled region and the unreshuffled one. In the reshuffled region ($k > 15$) we cannot find a clear tendency about five curves as in the energy transfer. Five curves are closer to each other in the larger scales ($k < 10$). The curves F and C are not so close as those in the energy transfer spectra. The small scale reshuffling does not affect the larger scales noticeably.

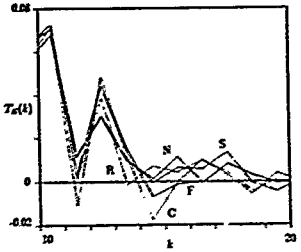


FIG. 3. Helicity transfer spectra in lin-lin scale for (a) RH. for $15 < k < 32$.

Finally we consider the helicity transfer spectra for (b) MH whose figure is also omitted. We point out one thing, that is, that the tendency is very similar to the energy transfer spectra although the maximum helicity condition: $T_H(k) = 2kT_E(k)$ is not satisfied.

IV. CONCLUDING REMARKS

We have compared the reshuffled fields, HR,HC and SA from the undisturbed one N and the filtered one F about both energy and helicity transfer spectra. It is found that in all cases the HC method is the most effective in destroying the ordering of the transfer of three and it works as if it were mode-filtering except the helicity transfer in random helicity forcing case. By keeping the correlation between the imaginary parts of all modes the energy is transferred in the HR. The correlation between modes is more important than that within each mode (i.e. the helicity associated phase) The SA method gives the results similar to the undisturbed cases This suggests that the *amplitude correlation* between the real part and the imaginary one may not be so essential to the nonlinear transfer as the phase relations. When we applied all three methods to the range, $15 < k < 32$ simultaneously, we obtained almost the same result as the case HC. Hence keeping the helicity spectrum has no special meaning. The helicity transfer in maximum helicity forcing case has characters similar to the energy transfer.

REFERENCES

[1] E. Levich, Y. Murakami and L. Shtilman, submitted to Phys. Lett. A.

[2] Y. Murakami, L. Shtilman and E. Levich, submitted to Phys. Fluids A.

[3] W. Polifke and L. Shtilman, Phys. Fluids, A1 (1989) 2025.

[4] W Polifke, Thesis, City College of CUNY, 1990.

[5] L Shtilman and W. Polifke, Phys Fluids A1, (1989) 778.

INFRARED AND ULTRAVIOLET BEHAVIOR OF TURBULENCE

TOHRU NAKANO
Department of Physics
Chuo University
Kasuga 1-13-27, Bunkyo-ku
Tokyo 112, Japan

Abstract In the present paper we discuss two ways of renormalization group method in turbulence: one by eliminating shorter wavelengths and the other by eliminating longer wavelengths. The former procedure gives the eddy viscosity and the additional correlation of external forces as a function of the cutoff wavenumber, predicting a reasonable value of Kolmogorov constant. The latter way helps us build a model, which may interpret the intermittent effects in fully developed turbulence.

I. INTRODUCTION

Since a renormalization group (RNG) method was successfully applied to critical phenomena by Wilson [1], turbulence has been one of the most attractive objects for the application of the method. The difficulty in the application to turbulence, however, seems to stem from the fact that the fluctuation source is located in low wavenumber region, while the sink is in high wavenumber region. Under this circumstance the simple elimination of high wavenumber components cannot renormalize the fluctuation source term appropriately and the elimination of low wavenumbers, on the other hand, fails to express the sink in a renormalized way. Hence, the simultaneous elimination of low and high wavelengths seems to be needed; every eddy receives energy mainly from larger eddies and transfers it to smaller eddies at the same time, which must be an essential cascade unit. Such a unit might not be established by the RNG method carried out in one direction.

In Sec.II we ask the infrared behavior of turbulence by eliminating shorter wavelengths, in which external driving forces are presumed. The estimated value of Kolmogorov constant and the coefficient of the eddy viscosity are in reasonable agreement with the current values. In Sec.III the ultraviolet behavior is studied by eliminating longer wavelengths. In the procedure the intermittency effects are examined in the framework of RNG.

II. INFRARED BEHAVIOR OF TURBULENCE

In turbulence the most contributions come from near eddies in wavenumber space. If we eliminate eddies with $q > \Lambda$, therefore, the eddy viscosity must be dependent on the cutoff Λ. MS [2] and ZVH [3] obtained such a cutoff dependent viscosity without relying on the distant interaction approximation in contrast to Yakhot and Orszag [4]. However, the additional stochastic forces cannot be forgotten in such a treatment. In the present paper, we numerically calculate the eddy viscosity $\nu(k,\Lambda)$ and correlation $D(k,\Lambda)$ of external forces as a function of Λ. To this end, we first assume that the bare external driving forces described by the correlation

$$< f_i(k,\omega) f_j(k',\omega') >= 2D_0 k^{-y}(2\pi)^{d+1}(\delta_{ij}-k_ik_j/k^2)\delta(k+k')\delta(\omega+\omega'),$$

as done by YO [4]. Eliminate the components with $\Lambda/s < q < \Lambda$ and rescale the velocity components to compare the original form. Several

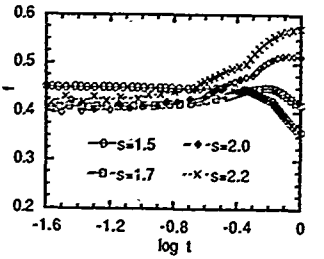


Fig.1. The eddy viscosity scaling function f for various values of s as a function of $t = k/\Lambda$.

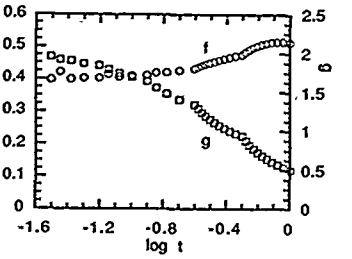


Fig.2. The scaling functions f and g for $s = 2$ as a function of $t = k/\Lambda$

values are chosen for s, ranging from 1.5 to 2.2. The obtained results slightly depend on the value. The recursion equations for the eddy viscosity and the correlation of forces are derived.

A. Eddy Viscosity and Correlation of External Forces

We express the viscosity and correlation in terms of scaling functions as $\nu(k,\Lambda) = k^{z-2}m(k/\Lambda)$ and $D(k,\Lambda) = k^{-y}n(k/\Lambda)$. Substitution of these expressions into the recursion equations, combined with the energy consideration, yields $y = d$ and $z = 2/3$ in accordance with the Kolmogorov scaling. Then, the viscosity and correlation take the forms $\nu(k,\Lambda) = \Lambda^{-4/3}f(k/\Lambda)$ and $D(k,\Lambda) = D_0k^{-d}(1+(k/\Lambda)^5 g(k/\Lambda))$. For the numerical calculation $D_0 = \pi^2$ has been employed. The scaling function $f(t)$ with $t = k/\Lambda$ is depicted for various s in Fig.1, where a cusp-like structure is seen near the cutoff for large s, in agreement with the observation by Kraichnan [5]. This is reasonable because he sets s equal to infinity, in principle, in the test-field model calculation. For small s such as 1.5 a cusp is not found, in accordance with MS [2] employing $s = 1/0.7 = 1.4$. ZVH[3] obtained the small cusp for the same value of s by introducing the triple nonlinear terms. Unfortunately, however, their triple terms are not Galilean-invariant. In Fig.2 we depicted $f(t)$ and $g(t)$ only for $s = 2$. We notice that the additional correlation of forces contributes to the original one by 50% near the cutoff, implying that quantitative corrections are significant as compared with the case in the absence of the additional force.. Another important point is that g acts as the backscatter from the subgrid scale in LES, reducing a true eddy viscosity considerably near the cutoff.

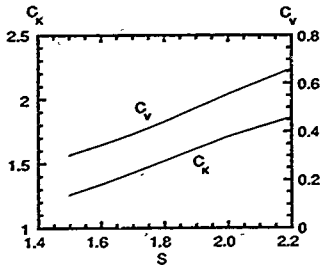


Fig.3. Kolmogorov constant C_K and the coefficient of the eddy viscosity C_V vs s.

B. Kolmogorov Constant

Put $E(k) = C_K\varepsilon^{2/3}k^{-5/3}$ and $\nu(k) = C_V\varepsilon^{1/3}k^{-4/3}$. The above numerical results, combined with the requirement $C_V/C_K^2 = 0.1904$ by the energy conservation, estimate C_K and C_V vs s, as shown in Fig.3. Both numbers are quite reasonable although they weakly depend on s. The details in this section will be reported elsewhere [6]

III. ULTRAVIOLET BEHAVIOR OF TURBULENCE

The purpose in this section is to investigate how the intermittency effects are appreciated in the framework of RNG. Since those effects are enhanced with decreasing scale, the successive elimination of longer wavelengths must reveal the existence of the intermittent effects. In order to proceed in that direction, we had better recall the experimental situation, in which pth order moments of velocity $v(r)$ on scale r are observed to be $< v(r)^p > \sim \varepsilon^{p/3}(L/r)^{\gamma_p} \sim r^{\zeta_p}$. $\zeta_p = p/3 - \gamma_p$ is not simply proportional to p. For small p ζ_p is almost $p/3$, but for large p it is far smaller than $p/3$. This fact means that a single scaling exponent for the velocity field cannot interpret the exponents associated with structure functions of all orders. It is hardly believed, therefore, that any simple modification of the exponent γ in the correlation of external forces will make satisfactory prediction on the intermittency effects. We need two scaling factors at least. A two-scaling model was proposed previously [7], in which ζ_p reads

$$\zeta_p = \begin{cases} p/3, & \text{if } p \le p_c \equiv 3z/(3z-2); \\ (1-z)p + z, & \text{if } p > p_c, \end{cases}$$

where z is a fitting parameter. If $z = 0.84$ is chosen ($p_c = 4.85$), the predicted values agree with the experimental data [8] quite well. Of interest is that the lower order moments obey the Kolmogorov scaling characterized by the exponent 1/3, while the higher order ones do the intermittent scaling characterized by the exponent $1 - z$.

The essential idea behind the RNG of elimination of longer wavelengths is as follows. To begin with, we have to presume a fictitious eddy viscosity, which drains energy from the system; molecular viscosity does not suffices. This fictitious viscosity is a counterpart of the presumed external forces in the process of elimination of shorter wavelengths. Consider the behavior of an eddy with wavenumber k, which interacts with eddies with $q < \Lambda$ and with $p \equiv |\mathbf{k} - \mathbf{q}|$. The contributions from the interaction are different, depending on whether p is smaller than Λ or not. (1) If $p < \Lambda$, the interaction terms act on the eddy k as stochastic forces. This case is possible only for $k < 2\Lambda$. The correlation of these forces is described by a type of the correlation assumed in Sec.II, although it is limited to the range $k < 2\Lambda$. The obtained correlation, combined with the presumed viscosity, derives the Kolmogorov scaling. (2) If $p > \Lambda$, the velocity components p must not be eliminated. The condition is realized for any value of k. The contributions in this case are conveniently decomposed into two parts: (a) the convection and (b) the distortion of the eddy k by larger eddies q. The convection velocity, which is predominantly determined by the largest eddies, carries all eddies together with no change in geometrical configuration; its effect is readily removed in the frame moving at the convection velocity. The distortion is, on the other hand, provided by the velocity gradient of larger eddies (mostly near eddies); the eddy is stretched in a certain direction (or directions) and shrinks it in another direction (or directions). The term completely different from the stochastic forces is expected.

The distortional contribution was investigated in the Fourier analized Navier- Stokes equation in Ref.[9], where we have shown that the strain rates due to eddies of a certain scale stretch and shrink the vorticity field of smaller scales, whose strain rates excite the vorticity of even smaller scales. If a chain of the excitation occurs randomly, the process may be treated as stochastic forces. If the excitation takes place systematically, however, a kind of coherent self-similar structure will be built in wavenumber space. A pair of oppositely-oriented vortex sheets are a candidate for such a self-similar structure; the two sheets approach each other and the vorticity is enhanced during the process. The time evolution of the vorticity, then, can be related to the dynamical scaling form proposed in Ref.[7]. Hence, the derived vortical structure is expected to be responsible for the intermittent effects. This picture will be developed in the framework of RNG

[1] K.G. Wilson, Phys. Rev. B4, 3184 (1971).

[2] V. Yakhot and S.A. Orszag, J. Sci. Comput. 1, 3 (1986).

[3] W.C. McComb and V Shanmugasundaram, J Phys A18, 2191 (1985).

[4] Ye. Zhou, G. Vahala and M. Hossain, Phys. Rev. A37, 2590 (1988).

[5] R.H. Kraichnan, J. Atmos. Sci. 33, 1521 (1976).

[6] T. Nakano, submitted to Phys. Fluids (1991).

[7] T. Nakano and M. Nelkin, Phys. Rev. A31, 1980 (1985).

[8] F. Anselmet, Y. Gagne, E.J. Hopfinger, and R.A. Antonia, J Fluid Mech. 140, 63 (1984).

[9] T. Nakano, Phys. Fluids, A2, 829 (1990).

FINITE ELEMENT MODELLING OF LOUDSPEAKER DRIVE UNITS

M A Jones and D J Henwood

Department of Mathematical Sciences/Information Technology Research Institute,
Brighton Polytechnic, Moulsecoomb, Brighton, U.K.
Also at B&W Loudspeakers Ltd, Steyning, W. Sussex, U.K.

ABSTRACT A Finite Element model for predicting the motion of axially-symmetric loudspeaker transducers using thin shell of revolution elements has been developed. The displacement of the diaphragm is calculated for a constant force, and the post-processing required to correct for the actual force is described. The effects of the electrical impedance of the voice coil and the mechanical impedance due to the motion of the cone are included.

1. INTRODUCTION

Although the ultimate evaluation of a loudspeaker depends on measuring the sound radiated by it, this radiated sound is not simply related to the vibration of the diaphragm caused by the signal fed to it. In order to apply the techniques of computer aided engineering to loudspeaker design, it therefore seems necessary to break the modelling process into stages.

The work presented here shows how a Finite Element (FE) model can be used to predict the vibration of a drive unit when it is driven by a constant force. We then describe a method of correcting for the real force provided by an amplifier. This approach allows the independent modelling of the electromagnetic parts of transduction. We present results for a model of a low frequency conical drive unit which can be compared with measured properties. Having gained confidence in the FE model, it is then possible to go on to predict the radiated sound pressure using methods such as the Boundary Element technique.

2. THE FINITE ELEMENT MODEL

The first stage in the calculation of the sound pressure produced by an arbitrary loudspeaker drive unit is to calculate the displacement of the diaphragm when it is driven by an arbitrary force. The motion of the diaphragm is governed by a partial differential equation and is therefore ideally suited to solution by the FE method.

Previous work (see for example Jones *et al* [1] and references contained therein) has assumed that the drive unit is axially symmetric. This assumption is reasonable both because the use of homogeneous, isotropic materials results in symmetric vibrations of the diaphragm, and because asymmetric modes are only weakly excited by the predominantly axially-directed force. The present computer code allows for the inclusion of asymmetries, thus making the model truly three-dimensional, but only at the cost of large losses in computational efficiency. In the following we shall restrict ourselves to describing the method used to model a conical drive unit for low frequency sound reproduction. The method is however applicable to other axially-symmetric structures and has been used successfully in the design of dome tweeters.

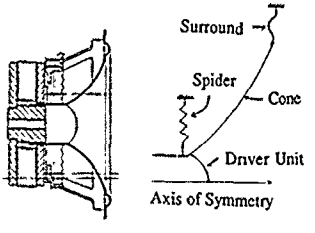


Figure 1 Schematic picture of a bass drive unit (left diagram) and the components of the idealisation used to model it.

A typical drive unit consists of the components shown in figure 1. The FE method requires values of Young's modulus, density, damping and Poisson's ratio for all the materials used in the structure in order to predict its displacement. A database has been set up which contains models of components already in construction and improved versions of these components. The accessing program uses a WIMP environment with one window leading on to another or to a specific I/O screen. This database easily permits alterations to components of the drive unit and the testing of novel design features.

The vibration of the structure is governed by the linear elasticity equation. Damping can be included in this equation by the use of a complex Young's modulus, $(1+j\eta)E$, in the stiffness matrix. The material loss factor η has been measured at B&W for particular materials [1]. It has been found that the drive unit can be discretised economically by the use of a three noded thin shell of revolution element as described in reference [1]. This particular element allows the displacement normal to the shell to vary as a fifth order polynomial instead of the more usual cubic. The element thus allows for internal bending to some extent.

If the force depends sinusoidally on time (i.e $F = f e^{j\omega t}$) then the displacements are similarly sinusoidal. When the structure is discretised the displacement of the system is defined by the solution of the set of linear equations

$$\left(-M\omega^2 + K\right) u = f \qquad (1)$$

where M and K are the global mass and stiffness matrices of the system, u is the complex vector of nodal displacements and f is the vector of applied forces.

A finite element program has been developed to form and solve the set of equations 1. The equations are solved using an efficient Gaussian elimination algorithm. An additional option is the use of modal analysis, giving information on the manner in which the cone breaks up at greatly reduced computational cost.

3. POST-PROCESSING

The Finite Element calculations described above eventually give the displacement vector of the cone for the constant force applied at the voice coil. The displacements calculated using this idealised force must be corrected for the effect of the real, frequency dependent force. In reality the current flowing in the coil is dependent both on the electrical circuit of which the coil forms a part and the equivalent electrical impedance of the mechanical parts, such as the cone itself. The force on the coil depends linearly on the current. Since the amplitude of the cone (and therefore its velocity) varies with frequency, the effect of the motion of the diaphragm must be included in the post-processing. The way in which this should be done was indicated by Shepherd and Alfredson [2].

Since the elasticity equation, eq. 1, assumes that the displacement of the diaphragm for a given frequency depends linearly on the force, the correction to the displacement is given by

$$u_{corr}(\omega) = u_{calc}(\omega) \cdot \frac{\tilde{F}(\omega)}{F} \qquad (2)$$

Shepherd and Alfredson define a common point, called the driving point, between the electrical and the mechanical parts of the transducer. This is the point at which the voice coil former joins the cone. Our finite element results calculate the frequency dependent velocity at this point, $v_D(\omega)$ $(= j\omega u_D)$, for a given constant force F. The total lumped mechanical impedance of the system at the driving point is defined by

$$Z_D = \frac{F}{v_D(\omega)} \qquad (3)$$

By calculating the current flowing in the voice coil due to the (constant) emf E supplied by the amplifier, we find by simple algebra that the force $\tilde{F}$ is given by

$$\tilde{F}(\omega) = E\left(\frac{BSZ_D}{Z_D Z_E + (BS)^2}\right) \qquad (4)$$

where BS is the magnetic coupling constant of the coil and $Z_E(\omega)$ is the impedance of the electrical circuit which can be modelled theoretically or empirically [3].

The correction to the calculated displacements, $\tilde{F}(\omega)/F$, adjusts both the amplitude and the phase of the results. The electrical circuit means that at high frequency the amplitude at a given point decreases linearly with ω. The mechanical parts damp the velocity when it is large, particularly at cone resonances.

4. COMPARISON OF RESULTS

The drive unit analysed here had a 13 cm diameter cone made of cobex and a 1.3 cm surround made of PVC. In the interests of clarity, results are shown for a unit without a dust cap.

The cone, surround and voice coil were discretised into 25 finite elements of the type described in section 2. The FE program was run and solutions obtained for the complex displacement of the diaphragm for a constant driving force applied at the voice coil along the axis of symmetry. The other boundary condition was that the surround was stationary at its edge. The FE results were then post-processed using the theory described in section 3, in order to correct for the real driving force supplied by the electrical circuit of the voice coil and the equivalent electrical circuit due to the finite impedance presented by the cone.

Figure 2 shows the FE results after post-processing. It can be seen that at low frequency the cone exhibits piston like behaviour, with a maximum in the velocity occurring towards the outer edge of the cone. The flat portion at the centre of the diagram corresponds to the zero velocity of the magnet pole piece. As the frequency increases the position of maximum velocity amplitude moves towards the centre, and beyond this point the cone starts to move out of phase with respect to its centre. We see a number of resonance peaks in the velocity of the whole cone, for example at frequencies of 1080 Hz, 2040 Hz, 2650 Hz, 3440 Hz and 4320 Hz. At about 1100 Hz the first breakup mode of the cone is seen, and above this frequency the number of separate parts of the cone vibrating out of phase with the voice coil increases.

The main differences between the pre- and post-processed data are found to be the magnitude of the maxima in the velocity, and the velocity of the voice coil at high frequency.

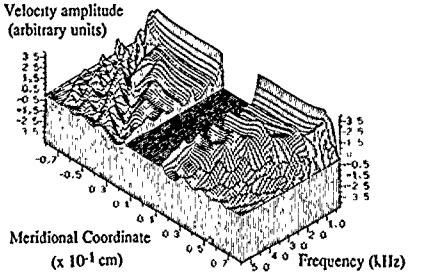


Figure 2 Finite Element prediction of the velocity of a drive unit along a diameter (meridian) of the diaphragm.

These results should be compared with laser Doppler interferometry measurements made at B&W Loudspeakers [4], shown in figure 3. The agreement between these sets of results appears to be good, the qualitative behaviour being similar and the maxima and minima in the velocity occurring at the same frequencies to within 10%. The measurements show a moderate amount of asymmetry in the behaviour of the drive unit which is not allowed for in the calculated results. The other main discrepancy is the velocity at the voice coil at high frequencies. This appears to be negligible in the measurements, but is seen to have a finite component from the calculations.

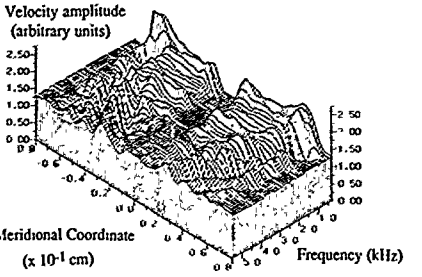


Figure 3 Measured velocity of a corresponding drive unit using laser Doppler interferometry.

5. CONCLUSION

We have described a method of predicting the vibration of axially-symmetric loudspeaker transducers using the FE method. The predictions have been compared with the velocities of existing drive units measured by a laser interferometer, and the model found to give substantially accurate results. The importance of this verification of the FE method is that if there are discrepancies between measured and predicted radiated sound pressures, then the errors must have been introduced by the method used to calculate the pressure.

Improvements to the calculation of the mechanical vibration of the diaphragm should take account of the interaction of the structure with the surrounding air. The construction of fully coupled structural-acoustic models is possible but at present they represent a significant increase in expense given the relatively small increase in accuracy gained. The development of faster algorithms for this modelling process is an area of current interest.

Areas requiring further work for a complete loudspeaker to be modelled accurately include modelling the magnetic field which governs the force on the diaphragm, modelling the interaction of the transducer with the cabinet and calculating the sound field created in a listening room.

ACKNOWLEDGEMENTS

The authors would like to thank B&W Loudspeakers Ltd, Steyning, and the SERC for their support of this project.

REFERENCES

1. C J C Jones, "Finite Element Analysis of Loudspeaker Diaphragm Vibration and Prediction of the Resulting Sound Radiation", PhD Thesis, Brighton Polytechnic (1985)

2. I C Shepherd and R J Alfredson, "An Improved Computer Model of Direct Radiator Loudspeakers", *J Audio Eng Soc* 33 (5) 322-329 (1985)

3. J R Wright, "An Empirical Model for Loudspeaker Motor Impedance", *J Audio Eng Soc* 38 (10) 749-754 (1990)

4. M A Jones, D J Henwood and P A Fryer, "Finite Element Prediction of the Vibration of Loudspeaker Diaphragms", to be published

NUMERICAL METHODS FOR USE IN MOLECULAR DYNAMICS SIMULATIONS

A.M.Mazzone
CNR - Istituto LAMEL, Via Castagnoli 1 - 40126 Bologna, Italy

Abstract - This work presents the algorithms that can be usefully implemented in molecular dynamics simulations and reports an analysis of the accuracy of such methods.

I. INTRODUCTION

Molecular dynamic simulation methods are currently used in condensed and solid state physics to evaluate many properties of liquids and crystalline targets.

The standard procedure upon which a molecular dynamic simulation is constructed is to consider a crystallite containing a number N of atoms interacting with realistic forces. The trajectories of the atoms, considered as point charges, are described by classical mechanics and are determined by solving the second order differential equation

$$\ddot{\mathbf{r}} = \mathbf{a}(\mathbf{r})$$

$$\mathbf{a}(\mathbf{r}) = \nabla V(\mathbf{r})$$

where r(t), v(t) and a(t) are vectors with 3N components, representing the x, y and z components of the position, velocity and acceleration of each of the N atoms in the computational cell. V(r) is the potential describing the interatomic forces.

Generally large N values, in the range 1000-10000, are used either to describe lattice properties in solid state theory or for applications in material processing, like ion implantation. Furthermore the characteristic times of the processes require from 10^4 up 105 integration steps. An additional difficulty arises from the strong non-linear character of the potential V(r).

This work revises the methods which can be advantageously used in molecular dynamics simulations and reports an analysis of the errors for various time-steps and cell dimensions.

II. METHODS

Owing to the duration of the simulations, multi-step methods which make use of the evaluation of a(r) at previous steps are to be preferred over those methods that make successive approximations to a(r) or that require the calculation of the derivative of a(r). On the other side the size of the vectors r, v and a render difficult the retention of information from many previous steps. A simple approach is the midpoint predictor

$$\mathbf{r}_{n+1} = \mathbf{r}_{n-1} + 2h\mathbf{v}_n$$

and the second-order Moulton corrector

$$\mathbf{v}_{n+1} = \mathbf{v}_c + (\mathbf{a}_{n+1} + \mathbf{a}_n)h/2$$

$$\mathbf{r}_{n+1} = \mathbf{r}_n + (\mathbf{v}_{n+1} + \mathbf{v}_n)h/2$$

where h represents the time step $h = t_{n+1} - t_n$. This method is referred in the Tables below as P(EC). This indicates that a predictor is used to calculate r and there are iterations of the sequence evaluation of a, correction of v and r. Also a third-order Adams-Bashforth predictor

$$\mathbf{r}_{n+1} = \mathbf{r}_n + (23\mathbf{v}_n - 16\mathbf{v}_{n-1} + 5\mathbf{v}_{n-2})h/12$$

and Adams-Moulton corrector can be used in this sequence

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (5\mathbf{a}_{n+1} + 8\mathbf{a}_n - \mathbf{a}_{n-1})h/12$$

$$\mathbf{r}_{n+1} = \mathbf{r}_n + (5\mathbf{v}_{n+1} + 8\mathbf{v}_n - \mathbf{v}_{n-1})h/12$$

These methods will be indicated as P(EC)AB-AM.

Alternatively Runge-Kutta (RK) methods can be applied which in the lowest order take the form

$$\mathbf{v}(t+h) = \mathbf{v}(t) + [2\mathbf{a}(t+h) + 5\mathbf{a}(t) - \mathbf{a}(t-h)]\,h/6$$

$$\mathbf{x}(t+h) = \mathbf{x}(t) + \mathbf{v}(t)h + [4\mathbf{a}(t) - \mathbf{a}(t-h)]\,h^2/6$$

III. RESULTS

A series of tests has been performed in order to compare the various methods described above. The simulation cell represents a silicon lattice with N in the range 1000-5000. The interatomic forces are described by a simple Morse-type potential of the form

$$V(r) = A\,[\exp-2a(r-r_e)-2\exp-a(r-r_e)]$$

The atomic coordinates and velocities at the beginning of the simulations are adjusted to correspond to the ones of a lattice at room temperature. As a test for numerical accuracy the energy conservation is used. In Table I we report the average energy change per step <dE/step> ([meV]) taken on 1000 steps for $h = 10^{-14}$ s In Table II a time-step $h = 10^{-13}$ s one order of magnitude larger has been used. For the iterative methods the number of iterations is equal to 2, as this value represents a good compromise between accuracy and computer times.

TABLE I — $h = 10^{-14}$ s

	N = 1000	N = 5000
method	<dE/step>	<dE/step>
P(EC)	~ 10^{-7}	~ 10^{-7}
P(EC)AM-AB	~ 10^{-5}	~ 5×10^{-5}
RK	~ 10^{-3}	~ 8×10^{-3}

TABLE II	$h = 10^{-13}$ s	
method	N = 1000 <dE/step>	N = 5000 <dE/step>
P(EC)	$\sim 10^{-7}$	$\sim 5 \times 10^{-6}$
P(EC)AM-AB	$\sim 10^{-5}$	$\sim 10^{-3}$
RK	$\sim 10^{-2}$	$\sim 5 \times 10^{-2}$

It is seen that

i) the simple P(EC) offers a stable and highly accurate solution.

ii) the use of P(EC)AM-AB leads to errors two orders of magnitude larger

iii) the worse results are obtained with RK. It has been found that the increase of the order of the RK does not lead to significative decrease of the errors.

In conclusion low order, iterative methods seems to be preferred for applications in molecular dynamics simulations.

A FAST METHOD FOR APPROXIMATE SEISMIC RAY-TRACING.

W.F.D. Theron
Department of Applied Mathematics
University of Stellenbosch, South Africa.

ABSTRACT

This paper describes a fast method for approximate seismic ray-tracing in a two-dimensional field, based on a generalisation of the method described by Vidale (1988) for finding the travel-time field from the finite difference approximation of the eikonal equation.

The method described here is eminently suitable for implementation on a local-memory message-passing multi-processor.

INTRODUCTION.

The work which is described here is the first stage of a program for seismic tomography, in which a large existing data-base of natural seismic events will be used to gain insight into the state of the rock mass. This software will be incorporated into the Integrated Seismic System [1].

The physical properties of the rock mass are specified by its slowness field, which is a scalar field with the slowness at a point defined as the inverse of the speed with which a P-wave propagates at that point. This slowness field is discretised, and the values at each grid point are the basic unknowns in our problem.

The tomographic method is described by Kissling [2], and is based on the idea that the travel-time for a very large number of seismic rays is known, where each ray travels along a path from the source of the event, through the slowness field, to the receiving station. An initial slowness field is assumed, and an approximate ray-tracing method is used to calculate the travel-time through this field for each ray. The difference between the measured and calculated travel-times is minimised by iteratively improving the slowness field. The coefficients of the equations used in this minimising procedure are, for each ray, the partial derivatives of the travel-time with respect to the slowness at each point.

Various methods have been described for carrying out approximate seismic ray-tracing. The method we have implemented makes use of the travel-time field, which is calculated as described by Vidale [3]. At this stage only the two-dimensional case has been implemented; we are busy extending to the three-dimensional case as described by Vidale [4].

Two schemes for tracing the rays through this travel-time field have been implemented.

Motivation.

We envisage applying this tomography program using the data of say one thousand events, measured at say 25 receiving stations, requiring the tracing of twenty-five thousand rays for each iteration of the method described above. The need for a very fast ray-tracer is self evident !

Parallelisation.

The method described here is eminently suitable for implementation on a local-memory message-passing multi-processor, as the rays corresponding to any given event are completely independent of those for the other events, so that the ray-tracing phase of the program is embarassingly parallel.

This is a typical "farm" type of parallel program, in which a driving program controls the execution, and the different worker tasks calculate the rays for each event and send the required results to the driver.

Furthermore, this is a very coarse grained algorithm, with far more calculation than communication, so that a high efficiency can be expected.

THE MATHEMATICAL MODEL.

Notation for the continuous model.

Given a rectangular region in the vertical x-z plane, we define the slowness at a point as the inverse of the speed of propagation of a P-wave at the point : $s(x,z) = 1/v(x,z)$.

The travel-time field for a particular event is denoted by $t(x,z)$, which is a scalar field of values of the time required for the wave to travel to the point (x,z).

Based on certain assumptions which will not be dealt with here, the relationship between the travel-time t and slowness s is given by the eikonal equation (1) :

$$(\partial t/\partial x)^2 + (\partial t/\partial z)^2 = s^2 \qquad (1)$$

The travel-time for ray number r is the time required to travel from the source to the station, and is denoted by T_r, where

$$T_r = \int s(x,z)\, dr, \qquad (2)$$

taking the line integral along the ray path.

Notation for the discrete model.

We discretize the region under consideration with horizontal dimension H*Nx, where Nx is the number of elements along the x-axis, and vertical dimension ζ*H*Nz, with ζ a dimensionless ratio of the sides of the rectangular elements. This is a slight generalisation of Vidale's approach, which is based on square elements.

S(i,j) and T(i,j) denote the slowness and travel-time respectively at the gridpoint positioned at depth (row) i and column j. The line integral (2) will be approximated by taking the sum of travel-times over short straight segments of the ray path.

The final results required of this program are the partial derivatives of the travel-time with respect to the slowness at each point :

$$D_r(i,j) = \frac{\partial T_r}{\partial S(i,j)} \qquad (3)$$

Calculating the travel-time field

The first step is to obtain values for the travel-time to each point, using Vidale's method for two dimensions [3]. This method can be summarised as follows :

Assume given values for S(i,j) and the location of the source of the event. The values of T(i,j) for the whole field are then calculated by starting at the source and extending the solution of T(i,j) in rings by using various finite difference approximations of the eikonal equation.

The detailed equations for the various cases are given in Appendix A. These reduce to the equations given by Vidale for the case $\zeta = 1$.

Approximate ray-tracing.

The next step is to use this travel-time field to trace the ray from the event to the receiving station, and to calculate the partial derivatives of travel time with respect to slowness at each point.

At any particular point on the ray, the gradient of the travel-time field can be determined, again using finite difference approximations. This gives a vector

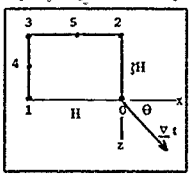
$$\text{grad}(t) = \nabla t = (\partial t/\partial x\ ;\ \partial t/\partial z),$$

the negative of which gives the direction from which the ray came.

Starting at the receiving station, we move backwards step by step to the point where the event occurred. The travel-time of a ray is found by accumulating ΔT_r, the increments for each

segment of the ray. The partial derivatives (3) can also be accumulated for those points affected by each segment of the ray.

In the initial implementation two schemes have been investigated. In the 3-point scheme the ray is forced to travel either horizontally, vertically or diagonally through the corners of a particular element, as follows :



Move to point 1 if $0 < \tan\theta < \frac{1}{2}\zeta$

$$\Delta T_r = H\,\tfrac{1}{2}(S_0 + S_1).$$

$$\partial T_r/\partial S_i = \tfrac{1}{2}H, \quad i = 0,1.$$

Move to point 2 if $2\zeta < \tan\theta < \frac{1}{2}\pi$

$$T_r = \zeta H\,\tfrac{1}{2}(S_0 + S_2).$$

$$\partial T_r/\partial S_i = \tfrac{1}{2}\zeta H, \quad i = 0,2.$$

Move to point 3 if $\frac{1}{2}\zeta < \tan\theta < 2\zeta$

$$T_r = H\sqrt{1+\zeta^2}\;\tfrac{1}{4}(S_0 + S_1 + S_2 + S_3).$$

$$\partial T_r/\partial S_i = \tfrac{1}{4}H\sqrt{1+\zeta^2}, \quad i = 0,1,2,3.$$

The more accurate 5-point scheme also allows the ray to go through the midpoints of the sides of the element. Equations similar to the above can easily be derived.

RESULTS.

The various algorithms have been implemented in a program written in 3L-FORTRAN and executing on a T800 transputer with 1 Mbyte of external memory. We intend implementing them in a parallel program.

Two results are of interest at this stage, namely results relating to accuracy of the calculated travel-time, and speed of execution. Both results are dependent on the shape of the region, the position of the source and the positions of the receivers, and the results given below for the specific example are preliminary results given solely for the purpose of getting an indication of the order of magnitude.

As our typical example we use the case of Nx = 2 Nz and ζ = 0.5, with the source in the bottom left corner and 25 stations round the edges opposite the source, for large grids (Nz > 50).

The total time to calculate the travel-time field for a grid with N points, N = Nx x Nz, plus 25 rays with their associated partial derivatives, was found to be of the order of 325 N μs using the 3-point scheme, and 362 N μs using the 3-point scheme.

Using a constant slowness value to obtain exact values for the travel-time, the maximum errors found for the calculated travel-time to the different stations were :

using the 3-point scheme : 2.8 %

using the 5-point scheme : 0.8 %.

The reduction in calculation time for obtaining the travel-time field when using the simplified equations for ζ = 1 was found to be 1% of the total time.

It is hoped to present detailed results of the parallel implementation at the conference.

REFERENCES.

[1] Mendecki, A.J. The Integrated Seismic System, Paper to be submitted.

[2] Kissling, E. (1988). Geotomography with local earthquake data, *Reviews of Geophysics* 26 659-698.

[3] Vidale, J.E. (1988). Finite-difference calculation of travel times. *Bull. Seism. Soc. Am.* 78, 2062-2076.

[4] Vidale, J.E. (1990). Finite-difference calculation of traveltimes in three dimensions. *Geophysics* 55, 521-526.

The work described herein was carried out as a contract between the Bureau of Industrial Mathematics of the University of Stellenbosch, and Advanced Mining Software.

APPENDIX A : FINITE DIFFERENCE EXPRESSIONS USED IN THE EIKONAL EQUATION.

An abreviated notation is used here. We write T_i for the travel-time to point i, i = 0,1,2,3,4 and S_i for the corresponding slowness values, with the points defined in the diagrams.

Three cases can be identified. In all cases, only the points in the first quadrant are shown in the diagrams. Identical equations are found for the remaining three quadrants.

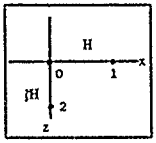
1. Starting at the source.

To start the solution, we take $T_0 = 0$ at the source, and obtain the travel-time to a neighbouring point 1 on the x-axis by taking the average value of the slowness between the two points :

$$T_1 = H\,\tfrac{1}{2}(S_0 + S_1).$$

Similarly, for point 2 in the z-direction

$$T_2 = \zeta H\,\tfrac{1}{2}(S_0 + S_2).$$

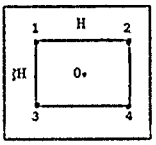


2. Cornerpoints.

With the travel-time known at three points (points 1,2,3) of a rectangular element, the time at point 4 can be found by writing the finite difference form of (1) for the center of the rectangle (point 0), as follows :

Take the average slowness at the centre as

$$S_0 = \tfrac{1}{4}(S_1 + S_2 + S_3 + S_4).$$



Define the parameters

$$\alpha = 1/(\zeta^{-2} + 1)$$

and

$$\beta = \alpha(\zeta^{-2} - 1).$$

Use central difference formulas for the centre of the rectangle, to obtain :

$$\partial t/\partial x = (1/H)\;[\,\tfrac{1}{2}(T_4 + T_2) - \tfrac{1}{2}(T_3 + T_1)\,]$$

$$= (1/2H)\;[(T_4 - T_1) + (T_2 - T_3)]$$

and

$$\partial t/\partial z = (1/2\zeta H)\;[(T_4 - T_1) - (T_2 - T_3)].$$

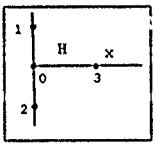
Using these expressions in the eikonal equation (1) we obtain a quadratic in $(T_4 - T_1)$, from which we can solve for T_4 :

$$T_4 = T_1 + \beta(T_3 - T_2) + \sqrt{\alpha(2H S_0)^2 - (1-\beta^2)(T_3 - T_2)^2}$$

For the special case with ζ = 1, α = 1/2, β = 0, and this reduces to the equation derived by Vidale (1988).

3. Midpoints.

With Ti known at 3 points (0,1,2) on a column, the travel-time to the mid-point on the next column can be found by using a forward difference formula in the x-direction, and central differences for the z-direction, at point 0, to obtain :



$$T_3 = T_0 + \sqrt{(H S_0)^2 - \tfrac{1}{4}(\zeta^{-2})(T_2 - T_1)^2}\}.$$

Similarly, with 3 points known on a row .

$$T_3 = T_0 + \zeta\sqrt{(H S_0)^2 - \tfrac{1}{4}(T_2 - T_1)^2}\}.$$

A MONTE CARLO SCHEME FOR CALCULATING NONLINEAR RADIATION TRANSPORT PROBLEMS

FENG TINGGUI
Institute of Applied Physics and Computational Mathematics, P.O. BOX 8009,
Beijing 100088, P.R.CHINA

Abstract

A Monte Carlo scheme for calculating nonlinear radiation transport problems is developed on the assumption that the thermal radiation emission is governed by local thermaldynamics equilibrum (LTE). Some numerical example is given and a conclusion is drawn from the numerical results.

I. INTRODUCTION

That thermal radiation interacts with matter in processes of its transfer forms radiation hydrodynamics problems. Thermal radiation here is taken to mean electromagnetic radiation of atomic origin, obtained from the processes of scattering, absorption, and thermal emission, the matter is generally modeled as a fluid of electrons and ions whose motion is governed by the equations of hydrodynamics The radiation field is described by a Boltzman transport equation for photons These equations are so complex that it is necessary to simplify the geometry or take some approximations of equations for solving them. The most widely used approximation to Boltzman equation is one group diffusion theory. But in some situations diffusion theory is inadequate, we need to solve directly the transport equation. Some available schemes has been given [1,2]. We here discuss a scheme for solving radiation transfer problems involving scattering process.

II. MATHEMATICAL DESCRIPTION OF THE PROBLEM

For simplify, some assumptions are taken as follows:

a. The electron temperature equilibrates with the ion, we call them material temperature.
b. The matter is stationary.
c. The thermal radiative emission is governed by LTE.
d. Thermal conduction in the matter is ignored.

With above assumptions, the radiation transfer problems is described as following equations [3].

$$\frac{1}{c}\frac{\partial I(\vec{r},\nu,\vec{\Omega},t)}{\partial t}+\vec{\Omega}\cdot\nabla I(\vec{r},\nu,\vec{\Omega},t)+\sigma_t(\vec{r},\nu,T,I)I(\vec{r},\nu,\vec{\Omega},t)$$
$$=\frac{c}{4\pi}\sigma_a(\vec{r},\nu,T)b(\nu,T)aT^4(\vec{r},t)$$
$$+\int_0^\infty d\nu'\int_{4\pi}d\Omega'\,I(\vec{r},\nu',\vec{\Omega}',t)\frac{\nu}{\nu'}\sigma_s(\vec{r},T,\nu'\to\nu,\vec{\Omega}'\cdot\vec{\Omega}), \qquad (1)$$

$$C_v\frac{\partial T(\vec{r},t)}{\partial t}=\int_0^\infty d\nu\,\sigma_a(\vec{r},\nu,T,I)\int_{4\pi}I(\vec{r},\nu,\vec{\Omega},t)d\Omega$$
$$-4aT^4(\vec{r},t)\int_0^\infty\sigma_a(\vec{r},\nu,T)b(\nu,T)d\nu$$
$$-\int_0^\infty d\nu\int_{4\pi}d\Omega\int_0^\infty d\nu'\int_{4\pi}d\vec{\Omega}'\sigma_s(\vec{r},T,\nu'\to\nu,\vec{\Omega}'\cdot\vec{\Omega})$$
$$\frac{\nu}{\nu'}I(\vec{r},\nu',\vec{\Omega}',t). \qquad (2)$$

$$\sigma_t(\vec{r},\nu,T,I)=\sigma_a(\vec{r},\nu,T)+\int_0^\infty d\nu'\int_{4\pi}d\Omega'\sigma_s(\vec{r},T,\nu\to\nu',\vec{\Omega}\cdot\vec{\Omega}')$$
$$+\int_0^\infty d\nu'\int_{4\pi}d\Omega'I(\vec{r},\nu',\vec{\Omega}',t)\frac{c^2}{2h\nu^2}\frac{1}{\nu'^2}\{\frac{1}{\nu'^2}\sigma_s(\vec{r},T,\nu\to\nu',\vec{\Omega}\cdot\vec{\Omega}')$$
$$-\frac{1}{\nu}\sigma_s(\vec{r},T,\nu'\to\nu,\vec{\Omega}'\cdot\vec{\Omega})\}. \qquad (3)$$

Here $I(\vec{r},\nu,\vec{\Omega},t)$ is the specific intensity of radiation, $T(\vec{r},t)$ is the material temperature, $b(\nu,T)$ is the normalized Plank spectrum, C_v is the specific heat, c is the light speed and the others have the physics meanings as usual.

III. SEMIIMPLICIT SCHEME

There are some couplings in equations (1) and (2) The Plank function emission source and cross sections are functions of material temperature. In addition, the total section σ_t depends on the intensity of radiation. These couplings can be divided into strong and weak categories [4]: Strong coupling refers to the temperature dependence of the emission source in EQ (1) and the radiation field dependence of the temperature in EQ (2); weak coupling refers to the temperature and intensity dependence of the cross section.

We divide time interval into some time steps: $\Delta t^{n+1/2}=t^{n+1}-t^n$, $0=t^1<t^2<\cdots<t^n<\cdots$, and then solve EQS (1) and (2) in each cycle Δt with strategy that treats strong coupling implicitly and weak coupling explicitly. Let

$$E(\vec{r},t)=aT^4(\vec{r},t). \qquad (4)$$

Substituting (4) for EQ (2) and using a backward Euler time differencing of EQ (2), we obtain

$$E^{n+1}=\Delta t^{n+\frac{1}{2}}\beta(\vec{r},T^{n+1})f(\vec{r},T^{n+1})$$
$$\int_0^\infty d\nu\,\sigma_a(\vec{r},\nu,T^{n+1},I^{n+1})\int_{4\pi}I^{n+1}(\vec{r},\nu,\vec{\Omega})d\Omega$$
$$-\int_0^\infty d\nu\int_{4\pi}d\Omega\int_0^\infty d\nu'\int_{4\pi}d\vec{\Omega}'\sigma_s(\vec{r},T^{n+1},\nu'\to\nu,\vec{\Omega}'\cdot\Omega)$$
$$\frac{\nu}{\nu'}I^{n+1}(r,\nu',\vec{\Omega}')\}+f(\vec{r},T^{n+1})E^n. \qquad (5)$$

where

$$\beta(\vec{r},T)=4aT^3(r,t)/c_v,$$
$$f(\vec{r},T)=1/(1+c\Delta t\beta(\vec{r},T)\sigma_a(\vec{r},T)),$$
$$\sigma_a(\vec{r},T)=\int_0^\infty\sigma_a(\vec{r},\nu,T)b(\nu,T)d\nu.$$

We use again a backward Euler time differencing of EQ (1) and substitute (5) for the term of emission source to obtain

$$\frac{1}{c}\frac{I^{n+1}-I^n}{\Delta t^{n+1/2}}+\Omega\cdot\nabla I^{n+1}+\sigma_t(\vec{r},\nu,T^{n+1},I^{n+1})I^{n+1}$$
$$=\frac{1}{4\pi}\chi(\vec{r},\nu,T^{n+1})\int_0^\infty d\nu\,\sigma_t(\vec{r},\nu,I^{n+1})$$

$$\int_{4\pi} I^{n+1}(\vec{r},\nu,\vec{\Omega})d\Omega' +$$

$$\int_0^{\infty} d\nu' \int_{4\pi} d\Omega' \sigma_s(\vec{r},T^{n+1},\nu' \to \nu\vec{\Omega}\cdot\vec{\Omega}')$$

$$\frac{\nu}{\nu'} I^{n+1}(\vec{r},\nu',\vec{\Omega}') +$$

$$\frac{1}{4\pi} c\chi(\vec{r},\nu,T^{n+1}) f(\vec{r},T^{n+1}) \sigma_p(\vec{r},T^{n+1}) E^n, \qquad (6)$$

where

$$\chi(\vec{r},\nu,T) = \sigma_a'(\vec{r},\nu,T)\beta(\nu,T) / \sigma_p(\vec{r},T),$$

$$\sigma_a'(\vec{r},\nu,T,I) = [1 - f(\vec{r},T)]\sigma_p(\vec{r},\nu,T,I) - \sigma_s(\vec{r},\nu,T),$$

$$\sigma_s(\vec{r},\nu,T) = \int_0^{\infty} d\nu' \int_{4\pi} d\Omega' \sigma_s(\vec{r},T,\nu \to \nu'\vec{\Omega}\cdot\vec{\Omega}')\frac{\nu'}{\nu}.$$

We rewrite EQ.(6) with differential quotient substituting for difference quotient to obtain

$$\frac{1}{c}\frac{\partial I(\vec{r},\nu,\vec{\Omega},t)}{\partial t} + \vec{\Omega}\cdot\nabla I(\vec{r},\nu,\vec{\Omega},t) + \sigma_s(\vec{r},\nu,T^{n+1})I(\vec{r},\nu,\vec{\Omega},t)$$

$$= \frac{1}{4\pi} c\chi(\vec{r},\nu,T^{n+1}) f(\vec{r},T^{n+1}) \sigma_p(\vec{r},T^{n+1}) E^n$$

$$+ \frac{1}{4\pi}\chi(\vec{r},\nu,T^{n+1}) \int_0^{\infty} d\nu' \sigma_s(\vec{r},\nu,T^{n+1}) \int_{4\pi} d\Omega' I(\vec{r},\nu\vec{\Omega}',t)$$

$$+ \int_0^{\infty} d\nu' \int_{4\pi} d\Omega' \sigma_s(\vec{r},T^{n+1},\nu' \to \nu,\vec{\Omega}\cdot\vec{\Omega}')$$

$$\frac{\nu}{\nu'} I(\vec{r},\nu'\vec{\Omega}',t), \qquad (7)$$

$$t^n < t < t^{n+1},$$

with initial condition

$$I(\vec{r},\nu,\vec{\Omega},t)|_{t=t^n} = I^n(\vec{r},\nu,\vec{\Omega}).$$

Here E^n,T^n and I^n can be obtained from last cycle. The values of $T^{n+1}$and $I^{n+1}$are to be obtained from letting $T^{n+1}=T^n$,$I^{n+1}=I^n$ or from suitable extropolation.

In every cycle $t^n<t\leq t^{n+1}$, Eqs. (1),(2) describing radiation transfer problems now turn into Eqs. (7),(5) and (8)

$$T^{n+1} = \left(\frac{E^{n+1}}{a}\right)^{\frac{1}{4}} \qquad (8)$$

We note that there are no strong couplings in Eqs (7),(5),(8) and they can be solved explicitly, i.e., if Eq. (7) is solved for I^{n+1}, then Eq. (5) explicitly gives E^{n+1},and then Eq. (8) explicitly gives T^{n+1}. So the key solving radiation transfer problem is solving Eq (7). However, Eq. (7) has been such a linear transport equation in each cycle Δt that can be solved easily by Monte Carlo method [1~2 or discrete ordinates (Sn) method [4].

IV NUMERICAL EXAMPLES

We coduider one dimensional slab 8 cm in thickness, heated by a 1 kev blackbody source at x=0. The slab is divided into three zones. $D_1=(x_0,x_1)$, $D_2=(x_1,x_2)$, $D_3=(x_2,x_3)$, $x_0=0$, $x_1=4$, $x_2=5$, $x_3=8$(cm). Macroscopic cross section of each zone are given by

$$\sigma_{ai}' = \frac{n_i}{\nu^3}(1 - e^{-\frac{\nu}{T}}) \qquad (9)$$

$$\sigma_{si} = 0, \qquad i = 1,2,3$$

$n_1=n_3=0.1$, $n_2=10$, $C_{v1}=C_{v3}=100$, $C_{v2}=10000$ $\Delta t=1\times 10^{-1}\mu s$ ($\mu s=10^{-6}$second). Numerical results are shown in fig1, which shows the material temperature distributions at different moment and compares the Monte Carlo temperature with that by difference (discrete ordinates) method.

The conclusion to be drawn from these numerical results is that the Monte Carlo numerical results are the same as that by the discrete ordinates method, the statistical fluctuation of the radiation intensity and material temperature is small and appears not to be propagating as cycle goes on. The scheme is suitable for solving radiation hydrodynamics equations.

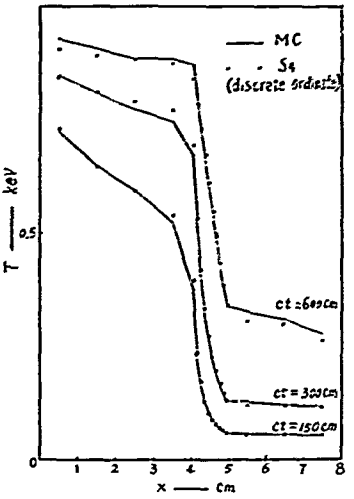


Fig 1. Temperature distributions calculated by MC and Sn.

REFERENCE

1. J A Fleck, JR., J.D Cummings, J. of Computation. Physics 8,313-342,1971.
2. J.A. Fleck, JR., and E.H. Canfield, J. of Computational physics 54, 508-523. 1984
3. Pomraning G C, The Equations of Radiation Hydrodynamics, Pergamon, Oxford, 1973.
4. R E. Alcouffe, B A. Clark, and E.W Larsen, in Multiple Time Scales, edited by J U. Brackbill and B Cohen (Academic Press Orlando, 1985), P.73

THE DYNAMIC EFFECT OF OROGRAPHY ON MESOSPHERE GRAVITY WAVES

Lai-Chen Chien
Institute of Physics
Academia Sinica
Nankang, Taiwan 11529
Republic of China

and

Eshung Wu
Department of Mathematics and Science Education
Provincial Taipei Teachers College
Taipei, Taiwan 10659
Republic of China

Abstract-The orographic effect on atmospheric gravity wave is investigated with two-dimensional, nonlinear, time dependent numerical simulation. The model of anelastic equations with heigth dependent basic potential temperature is applied to study the problem. The orographically generated gravity waves propagated upward to the middle atmosphere. The computation results show that wave properties agree with the existing solution.

I. INTRODUCTION

Gravity waves play an extremely important role in affecting the density perturbation in the middle and upper atmosphere (Hung and Lee, 1990). The investigation on the middle and upper atmospheric density changes is not only need for space vehicle design but also for numerical weather prediction. The projects such as the Space Shuttle, National Aerospace Plane, Space Telescope and Tethered Satellite benifit from such studies.

Numerous research efforts over the last decades have shown gravity wave motions play a significant-role in determining the circulation and structure of the middle atmosphere. Theoretical and observational studies demonstrate that gravity waves are able to transport momentum and energy over considerable distance. Lilly and Klemp (1979) investigated the orographic effect on the gravity wave transportation and discovered that the induced gravity wave transport the momentum and energy to meso-sphere. The energy observed and dissipated in the middle atmosphere. The gravity waves give rise to wave drage.

Smith (1979), Pitt and Lyons (1989) discovered that when the width of orography is greater than the induced gravity wave length, the gravity wave propagated upward and tilted upwindward. There exists a maximum wind velocity region in the orographic leeward. The linear theory explains that the gravity wave transmits upward, reflects from traupopause and transport the energy to accelerate the air at lee side.

Klemp and Lilly (1975), Pelties and Clark (1979) applied the non-linear theory to investigate the orographic effect on gravity wave. They indicated that the energy accelerating the lee side air comes from the dispersion of gravity waves. The theories on the mechanism between the gravity wave and 1 atmosphic motion are different based on various assumption.

The study of Lilly and Klemp (1979) indicated that the orographic effect on the gravity wave propagation plays an important role in atmospheric motion. Especially gravity waves induced by irregular orography, propagates upward high enough. Then the wave dissipates and the energy is absorbed by atmosphere. Therefore, the mechanism between gravity wave and meso-sphere atmospheric motion is widely investigated. The gravity wave saturation is the source of kinetic energy of atmospheric motion, e.g. the traupopsphere and stratosphere atmospheric motion.

The drag have parameterized by Canadian Climate Center (McFarlance, 1987) to investigate FCMWF numerical weather prediction phenomena. Wallace et. al. (1983) incorporated the parameter in studying North Hemisphere circulation during winter and reduced the medium range forecast error. Alpert, et. al. (1988) applied the balance method, and considered the orographic effect induced gravity wave drag, improved the medium range forecast accuracy in National Meteorological Center operation. Iwasaki, Yamada and Tada (1988) employed the nonhydrostatic gravity wave mode and investigated the gravity wave drag parameter for medium range forecast.

In this study, the anelastic model with heigth dependent basic potential temperature model developed by Bacmeister and Schoeberl (1989) is applied to study the orographic effect on the gravity waves. Applying the model described and initial-boundary value specified, we can solve the governing equations and obtain the gravity wave properties in the atmosphere.

II. MODEL DESCRIPTION

In this paper, the generation, development and damping of gravity waves induced by Orography is investigated. The model equations based on the modified form of the anelastic equations with height dependent basic potential temperation (Macmeister and Schoeberl, 1989) is incorporated. Using the primitive variables (u, v, p), we have the two dimensional momentum equation in horizontal (x) and vertical (z) component

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} = -\frac{\partial p^*}{\partial x} + F(x) \qquad (1)$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + w\frac{\partial w}{\partial z} = \frac{\partial p^*}{\partial z} + g\frac{\theta'}{\theta} + F(z) \qquad (2)$$

In order to couple density and velocity and to filter the acoustic wave, the local variation of density in equation of continuity must be neglected (Wilhelmsson and Ogura, 1972)

$$\frac{\partial \rho u}{\partial x} + \frac{\partial \rho w}{\partial z} = 0 \qquad (3)$$

And the energy equation is of the form

$$\frac{\partial \theta'}{\partial t} + u\frac{\partial \theta'}{\partial x} + v\frac{\partial \theta'}{\partial z} + w\frac{\partial \theta}{\partial z} = F(\theta'); \qquad (4)$$

In above equations, $\rho(z)$ is the density, θ, potential temperature; R, gas constant; P_{00}, the reference pressure.

The potential temperature can be split into the upwind steady potential temperature $\theta(z)$ and disturbance $\theta'(x, z, t)$,

$$\theta(x, z, t) = \theta(z) + \theta'(x, z, t) \qquad (5)$$

The diffusion damping terms in momentum equation (1) and (2) are

$$F(x) = k_2 \frac{\partial^2 u}{\partial x^2} + \frac{\partial}{\partial z}[k_2 + k_{turb}\frac{\partial u}{\partial z}]$$

$$+ k_4 \frac{\partial^4 u}{\partial x^4} - \gamma(x, z)[u - u_{upst}(z)], \quad (6)$$

$$F(z) = k_2 \frac{\partial^2 w}{\partial x^2} + \frac{\partial}{\partial z}[k_2 + k_{turb}\frac{\partial w}{\partial z}]$$

$$+ k_4 \frac{\partial^4 w}{\partial x^4} - \gamma(x, z)w, \quad (7)$$

where $\gamma(x, z)$ is Rayleigh damping coefficient. The diffusion damping term in energy equation (4) is

$$F(\theta') = k_2 \frac{\partial^2 \theta'}{\partial x^2} + \frac{\partial}{\partial z}[k_2 + k_{turb}\frac{\partial \theta'}{\partial z}]$$

$$+ k_4 \frac{\partial^4 \theta'}{\partial x^4} - \gamma(x, z)\theta' \quad (8)$$

The turbulent diffusion k_{turb} (Orlanski and Ross 1973) is

$$k_{turb} = \begin{cases} 0.1k_2 \dfrac{\Delta z}{\Delta z_0}\left[\dfrac{\Delta(\theta + \theta')}{\theta + \theta'}\right]^{1/3} & \text{when } \dfrac{\partial \theta'}{\partial z} < 0 \\ 0 & \text{when } \dfrac{\partial \theta'}{\partial z} \geq 0 \end{cases} \quad (9)$$

In equation (9), Δz_0 is 1 km (Bacheister and Schoeberl, 1989), Δz is the vertical grid distance. k_2 and k_4 in above equations are basic diffusion coefficient $k_2 = 10^{-4}m^2/sec$ and biharmenic diffusion coefficient $k_4 = 10^3/N$, where N is Brunt-Saeisaelae frequency

$$N^2 = \frac{g}{\theta}\frac{\partial \theta}{\partial z} \quad (10)$$

In order to consider the orographic effect, the coordinate transformation is applied

$$x^* = x, \quad z^* = z_0 \frac{z - h(x)}{z_0 - h(x)} \quad (11)$$

where h(x) is the orographic height and z_0, the characteristic height considered. Applying the coordinate transformation, we can transform the physical plane into Cartesian coordinate in computation plane.

Solution of the governing equations is initial-boundary value problem. The initial values of the atmospheric variables are those of background undisturbed data. Incorporating the periodic wave model developed by Bacheister and Schoebel (1989), we consider the wave motion is periodic and assume that the computation domain is periodic. The center of the computation domain in x-direction is the hightest point of the orography. The domain extends it range of length w in both directions. The sponge damping boundary conditions are applied (Figure 1).

The boundary condtions for the upwind are the undisturbed wind and potential temperature distribution,

$$u(z) = u_{upst}(z) \quad (12a)$$

$$\theta(z) = \theta(z) = \theta\exp(z/H_\theta) \quad (12b)$$

$$\text{i.e. } \frac{g}{\theta}\frac{\partial \theta}{\partial z} = \frac{g}{H_\theta} = N^2 \quad (12c)$$

where H_θ is scale height potential temperature. In general case, H_θ = 100 km and N = 0.01/sec. The upwind wind profile and temperature distribution are the sounding data. For the downstream, the sponge boundary conditions are applied. Because the viscosity effect are not considered, the free slip boundary condition is employed.

The initial conditions for the field considered are those of undisturbed conditions. At the very beginning of the time, the flow properties over the domain are uniform. Then the flow field is disturbed. The gravity wave is induced by the orographic effect and propogates.

III. RESULTS AND DISCUSSION

At first, the physical plane is mapped into computation plane by equation (11). Casting the governing equations into finite difference form and applying the alternating direction implicit algorithm, we obtain the numerical solution for the flow properties.

In order to check the accuracy of the model and the computer coding, we test the case of flow over a mountain. The orography is of bell shape with height 500 m, width 24 km. Assume the background undisturbed wind velocity is 15 m/sec. The disturbance appears because of the existence of the orography. At the initial time, t = 40 minutes, the disturbance propagates upward to z = 4 km (Figure 2). Figure 3 shows that the streamlines become steeper near the lee side. At time t = 80 minutes, there appears a periodic wave at height 3 km (Figure 4). At time t = 100 minutes, the flow pattern shows that the motion is dominated by the orography effect for the lower atmosphere, $z \leq 3$ km, whilst at upper atmosphere, it is controlled by the wave motion, Figure 5. As time goes on, t = 120, 140 minutes, because of the nonlinear mechanism, the dispersive phenomena appear (Figure 6, 7). The physical phenomena agree with these of investigation of Peltier and Clark (1979).

Then, we apply the program developed, to study the gravity wave motion in the atmosphere. Assume the thickness of the atmosphere be 100 km, and initial velocity distribution is function of height, equation (12a). Fritts and Dunkerton (1984) proposed the velocity distribution be of the form

$$u(z, 0) = -U \tanh[(z - z_c)/d] \quad (13)$$

where U = 30 m/sec, z_c = 60 km, d = 10 km, The horizontal wavelengh is 50 km. The vertical velocity disturbance is

$$w(x, t) = w_0 f(t) \sin kx, \quad (14)$$

the amplitude is function of time,

$$f(t) = \begin{cases} \sin^2(\pi t/t_f) & t \leq t_f \quad (15a) \\ 0 & t = t_f \quad (15b) \end{cases}$$

where t_f = 2, 3 or 4T, T is the period. The investigation data of Fritts (1985) is applied for u_0 = 1.0, 1.4, 2.0 m/sec. And horizontal wavelengths are 16.7, 25 , 50 km.

Similar to the previous case, uniform flow over orography, the disturbance is induced by the lower boundary orography. The instigated wave propagated upward as time goes on.

At first, we examine the time variation of potential temperature distribution. The disturbance (14) is added at initial time, $t \leq 2T$. Then, no additional disturbance further. As t = 3.0T and the lower atmosphere, $z \leq 50$ km, the potential temperature distribution is almost dominate by the background uniform flow, Figure 8a. The gravity wave spreads upward at 10 km < z < 60 km, the wave patterns are discernible. When t = 3.5T, Figure 8b, the rear wave moves faster than the front one. The wave crest and trough locate almost at vertical line. Figure 8c shows that at t = 4.0T. above 40 km, especially $z \leq$ 60km, the tendency of dispersive appears for the gravity wave development.

The disturbed horizontal velocity is also shown in Figure 9. At time 3.5 period, there arises unsteady phenomena. The disturbance velocity is greater than 10 m/sec. We assume that the uniform background velocity is of magnitude 30 m/sec. The disturbance grows up to one third of the background value, Figure 9a. For time goes to 4T, the disturbance increases to 20 m/sec at height 40 km, Figure 9b and 9c.

The investigation indicates that wave of λ = 50 km under the disturbance such as orography, behaves dispersive tendency. The phenomena agree with those investigated by Holten and Wehrbein (1980).

IV. CONCLUSION AND RECOMMENDATION

Gravity waves plays an important role in affection the density disturbance in the middle and upper atmosphere. In this investigation, model of modified form of the anelastic equation with height dependent basic potential temperature is applied to study the disturbance generated by orography. The perturbation induced by orography begins to display the dispersive phenomena. The relationship between the nonlinear effect and dispersion is obvious. The results agree with the study of Holten and Webrbein (1980)

According to the observation of VHF radar, the maximum amplitude of density perturbation caused by gravity waves associated with Typhoon at West Pacific Region is 15%. While for tropical storms at Northern America, the amplitude disturbance is ± 12% (Hung and Lee, 1990). This is an intresting problem. Besides the density variation with height, the gravity force modification must be considered in the further study.

REFERENCES

Alpert, J. C., Kanamitsu, M. and Caplan, P. M. (1988). Mountain Induced Gravity Wave Drag Parameterization in the NMC Medium Range Forecast Model. Pro. 8th Conf. Num. Weath. Pred., pp. 726-733.

Bacmeister, J. T. and Schoeberl, M. R. (1989). Breakdown of Vertically Propagating 2-D Gravity Waves Forced by Orography. J. Atm. Sci., Vol. 46, pp. 2109-2134.

Fritts, D. C. (1985). A Numerical Study of Gravity Wave Saturation: Nonlinear and Multiple Wave Effects. J. Atm. Sci. Vol. 42, pp. 2043-2057.

Holton, J. R. and W. M. Wehrbein (1980). A Numerical Model of the Zonal Mean Circulation of the Middle Atmosphere. Pure Appl. Geophys., Vol. 118, pp. 284-305.

Hung, R. J. and Lee, C. C. (1990). Atmospheric Density Remote Sensing of Mesosphere and Thermosphere to Be Used for Spacecraft Design by Adopting VHF Radar and HF Doppler Sounder at Low Latitude West Pacific Site during Winter. Acta Astronautica, Vol. 21, pp. 583-597.

Iwasak, T., Yamada, S. and Tada, K. (1988). Impact of a Parameterization Scheme for Orographic Gravity Wave Drag with Two different Vertical Partitions on MRF. Pro. 8th Num. Weath. Pred., pp. 734-740.

Lilly, D. K. and Klemp, J. B. (1979). The Effects of Terrain Shape on Nonlinear Hydrostatic Mountain Wave. J. Fluid Mech., Vol. 95, pp. 54-61.

McFarlane, N. A. (1987). The Effect of Orographically Excited Gravity Wave Drag on the General Circulation of the Lower Stratosphere and Troposphere. J. Atm. Sci., Vol. 44, pp. 1775-1800.

Orlanski, I. and Ross, B. B. (1973). Numerical Simulation of the Generation and Breaking in Internal Gravity Waves, J. Geophys. Res., Vol. 36, pp. 8808-8811.

Peltier, W. R. and Clark, T. L. (1979). The Evolution and Stability of Finite-Amplitude Mountain Waves. J. Atmos. Sci., Vol. 36, pp. 1498-1529.

Peltier, W. R. and Clark, T. L. (1980). Surface Wave Drag and Severe Downslope Windstorms, J. Atmos. Sci., Vol. 37, pp. 2122-2125.

Pitts, R. O. and Lyons, T. J. (1989b). Airflow over 2-D Escarpment. II. Theory. Q. J. R. Met. Soc., Vol. 115, pp. 982-995.

Smith, R. B. (1979). The Influence of Mountain on the Atmosphere. Adv. Geophys., Vol. 21, pp. 87-230.

Wallace, J. M., Simmons, A. J., Branstator, G. W. (1983). Barotropic Wave Propagation and Instability, and Atmospheric Teleconnection Patterns. J. Atm. Sci., Vol. 40, pp. 1363-1392.

Wilhelmson, R. and Ogura, Y. (1972). The Pressure Perturbation and the Numerical Modeling of a Cloud. J. Atm. Sci., Vol. 29, pp. 1295-1307.

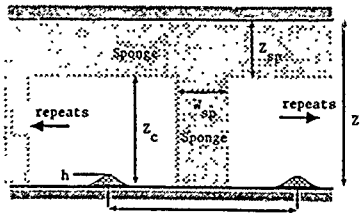


Figure 1. Computation domain

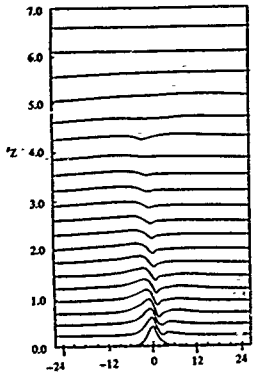


Figure 2. The flow pattern at t = 40 minutes for uniform flow over orography of height 500 m, width 12 km.

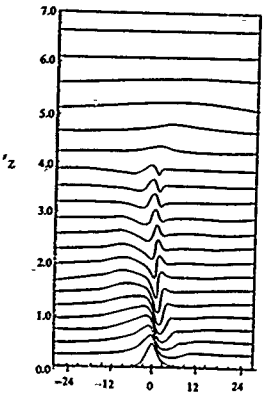


Figure 4. Flow pattern at t = 80 minutes for Figure 2.

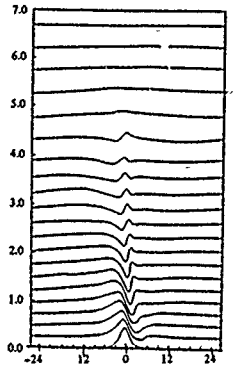


Figure 3. Flow pattern at t = 60 minutes for Figure 2.

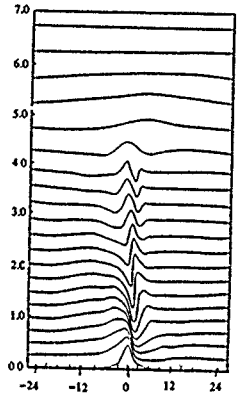


Figure 5. Flow pattern at t = 100 minutes for Figure 2.

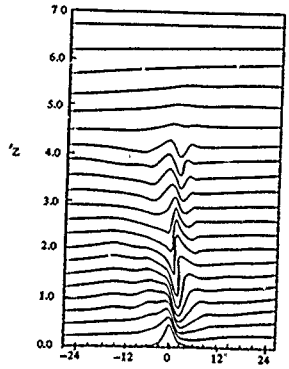


Figure 6. Flow pattern at t = 120 minutes for Figure 2.

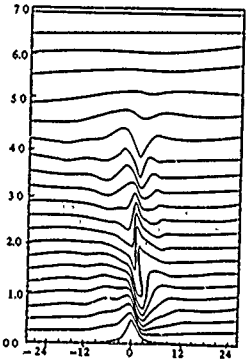


Figure 7. Flow pattern at t = 140 minutes for Figure 2.

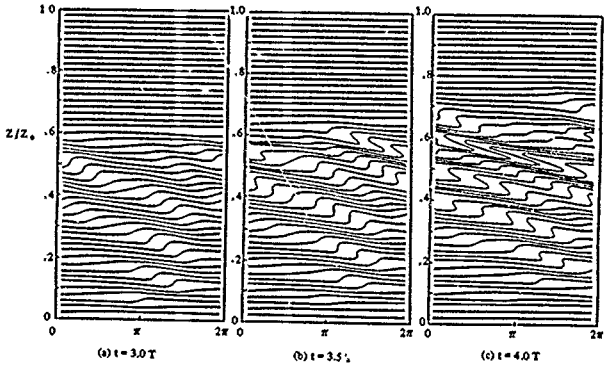


Figure 8. Potential temperature disturbance papagation at (a) t = 3.0T, (b) t = 3.5T, (c) t = 4.0T.

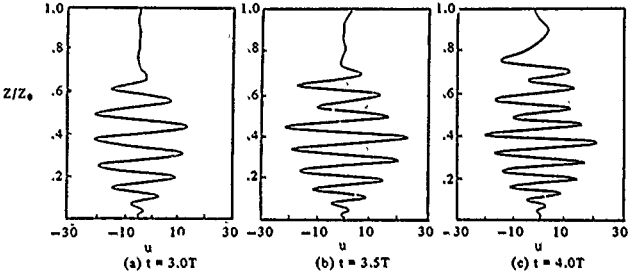


Figure 9. Time variation of horizontal velocity for wave disturbance at (a) t = 3.0T, (b) t = 3.5T, (c) t = 4.0T

NUMERICAL METHODS OF SOLVING BIHARMONIC-BOUNDARY VALUE PROBLEMS.

TIPARATANA WONGCHAROEN
Bangkok University, 12120
Pathumthanee Province, Thailand

AND

NGAMNIT WONGJAREON
Kingmongkut's Institute of
Technology Thonburi, Bangkok,
Thailand. 10140.

Abstract - In this work a system for calculate the solution of biharmonic equation is developed. In the main step of calculation by numerical method, there are n system equations and n variables being generted. The step of calculation are very cumbersome and tedious jobs. Program for calculate these solutions is written for the purpose of handling such problems. This technique is suitable as a simple and time saving. This program is used to calculate the solutions of biharmonic equation and display the results by printer and plotter.

I. INTRODUCTION

In this paper our object is to develop a finite-difference method to solve the biharmonic equation.

$$\frac{\partial^4 \phi}{\partial x^4} + 2\frac{\partial^4 \phi}{\partial x^2 \partial y^2} + \frac{\partial^4 \phi}{\partial y^4} = 0 \qquad (1)$$

This equation may by satisfied by taking the function ϕ in the following form:

$$\phi = \cos\frac{m\pi x}{l} f(y) \qquad (2)$$

where ϕ is the stress function. m is an integer. l is the length of rectangular beams. f(y) is a function of y only. Substituting equation (2) into equation (1) and using the notation $m\pi/l = \alpha$, we find the following equation for determining f(y):

$$f^{iv}(y) - 2\alpha^2 f^{ii}(y) + \alpha^4 f(y) = 0 \qquad (3)$$

The equation (3) can solve by finite-difference method.

II. METHODS

Finite-Difference method.

To solve a boundary-value problem by the method of finite differences, every derivative appearing in the equation, as well as in the boundary conditions, is replaced by an appropriate difference approximation. Central differences are usually preferred because they lead to greater accuracy. Using the difference approximation with $y_0 = 0$, $y_N = L$, and $Nh = L$, we will have

$$f_{n-2} + (-4 - 2\alpha^2 h^2) f_{n-1} + (6 + \alpha^4 h^4 + 4\alpha^2 h^2) f_n + (-4 - 2\alpha^2 h^2) f_{n+1} + f_{n+2} = 0, \quad n = 1, 2, \ldots, N-1 \qquad (4)$$

$$\begin{aligned} &\text{For } y = 0, \ f_0 = f(y_0) = A \\ &\text{For } y = N, \ f_N = f(y_N) = A \\ &\text{and } f_{-1} = f_1, \ f_{N+1} = -f_{N-1} \end{aligned} \qquad (5)$$

writing out (4) for n = 1, 2, ..., N-1 and using the boundary condition (5), we obtain the matrix which nonzero elements appearing only along the principal five diagonals. An algorithm is again available for solving directly such systems.

Solution of five-diagonal systems.

The algorithm consists in applying the following steps:

1. Compute the initial values

$$W_1 = C_1$$
$$\beta_1 = D_1 W_1, \ \beta_0 = 0, \ \beta_{N-1} = 0$$
$$\gamma_1 = E_1 W_1, \ \gamma_0 = 0, \ \gamma_{N-1} = \gamma_{N-2} = 0$$

2. Compute recursively

$$\delta_n = B_n - A_n \beta_{n-2}$$
$$W_n = C_n - A_n \gamma_{n-2} - \delta_n \beta_{n-1}$$

$$\gamma_n = \frac{D_n - \delta_n \gamma_{n-1}}{W_n}$$
$$\gamma_n = \frac{E_n}{W_n}$$

where $n = 2, 3, \ldots, N-1$

3. Compute

$$h_0 = 0$$
$$h_1 = \frac{b_1}{W_1}$$
$$h_n = \frac{b_n - A_n h_{n-2} - \delta_n h_{n-1}}{W_n}$$

where $n = 2, 3, \ldots, N-1$

4. Compute the values of f backward, using

$$f_{N-1} = h_{N-1}$$
$$f_n = h_n - \beta_n f_{n+1} - \gamma_n f_{N+2}$$

where $n = N-2, N-3, \ldots, 1$

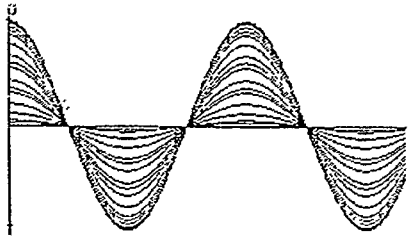
We can now apply these algorithms to obtain the solutions of biharmonics equation.

Computer results for biharmonic solution.

From article II, we obtain the solution of biharmenic equation by substituting the following conditions:

$$y_0 = 0 \ , \ y_N = 2.0$$
$$N = 20 \ , \ h = 0.1$$
$$f_0 = 150 \ , \ f_N = 150$$

The solution of biharmonic equation was programmed for computer



ACKNOWLEDGMENT

The authors thank Dr. Wudhibhan Prachyabrued for his suggestions which greatly improved the paper.

REFERENCES

1. J.P. Coleman, A new fourth-order method for y" = g(x)y + r(x), Preprint, Dept. of Mathematics, Univ. of Durham, England, March 1979.

2. V.I. Krylov and L.T. Shul'gina, Handbook of Numerical Integration [in Russian], Nauka, Moscow (1966), p.370.

ON THE NUMERICAL SOLUTION OF A MODEL FOR A SINGLE SPECIES POPULATION DYNAMICS

ASGHAR KERAYECHIAN
Department of Mathematics
Ferdowsi University of Mashhad
Mashhad, Iran

Abstract- In this paper we present a numerical method for solving an integro-differential equation which is a model for age-dependent populations with spatial diffusion. Our numerical scheme is based on finite difference method and discretization of the model yields a linear system with block tridiagonal matrix. The stability of the method is discussed by considering the eigenvalues of this matrix.

1. INTRODUCTION

There has been much work on mathematical theories of biological populations. The best known model of age-dependent population dynamics was first introduced by Von Foerster [1]. In this model it is assumed that the death and birth processes depend only on age. Several authors including MacCamy and Gurtin [2], Hoppenstead [3], Swick [4], drived a model based on the Von Foerster model in which the birth and death processes are allowed to depend, in addition to age, on total population size.

The problem of spatially nonhomogeneous, age-dependent population dynamics, i.e. the case where population density can vary with space, age and time has been considered by Zachmann and Logan [5] and MacCamy [6] and Gurtin [7].

The model in [6] concerns a single species population moving in a limited one-dimensional environment and avoiding crowding. The model is based on the following equations:

$$\rho_t+\rho_a=-q_x(x,a,t)-\lambda(a,P)\rho, \qquad 0<x<d,\ 0<t,\ 0<a\le L \tag{1.1}$$

$$P(x,t)=\int_0^L \rho(x,a,t)da \qquad 0<x<d,\ 0\le t \tag{1.2}$$

$$\rho(x,0,t)=b(x,t) \qquad 0<x<d,\ 0<t \tag{1.3}$$

$$\rho(x,a,0)=\rho(x,a) \qquad 0<x<d,\ 0\le a\le L \tag{1.4}$$

$$\rho(0,a,t)=\rho(d,a,t)=0 \qquad 0\le t,\ 0\le a\le L \tag{1.5}$$

where

$\rho(x,a,t)$ is the population density, that is, the number of individuals, per unit volume of age a at time t and position x, where x is restricted to an interval $0<x<d$.

$P(x,t)$ is the total population density and L is the maximum life span of individuals.

$q(x,a,t)$ represent the population flux and it is assumed to be proportional to the gradient of total population:

$$q(x,a,t)=-k(a)\nabla P(x,t) \tag{1.6}$$

where $k(a)>0$ is the diffusivity. The negative sign indicates that the flow of population always lies in the direction of decreasing density.

$\lambda(a,P)\ge 0$ is the death rate and the term $\lambda(a,P)\rho$ in Eqn.(1.1) is the loss of individuals of age a at x, due to deaths.

$b(x,t)$ is the birth process and is given by the regeneration rule of the form:

$$b(x,t)=\int_0^L \beta(a,P)\rho(x,a,t)da \tag{1.7}$$

where $\beta(a,P)$ is a non-negative function often called the birth-rate.

$\rho(x,a)$ is the initial age-space distribution.

The boundary conditions in Eqn.(1.5), biologically, assert that all individuals reaching a boundary leave the interval (0,d).

In [6], [7] the authors assume that the individuals can live to an arbitary age, so they take $L=+\infty$ in Eqns.(1.2) and (1.7). As was asserted in [6] and [7], finding a solution $\rho(x,a,t)$ of Eqns.(1.1)-(1.5) in general, is difficult and it was left as an open problem.

The purpose of this paper is to develop a numerical scheme to approximate $\rho(x,a,t)$ by simplifying assumptions as follows:

It will be assumed that k, the diffusivity, λ, the death rate and β, the birth rate are constants. Thus the mathematical model for $\rho(x,a,t)$ is:

$$\rho_t+\rho_a=k\int_0^L \rho_{xx}(x,a,t)da-\lambda\rho, \qquad 0<x<d,\ 0<t,\ a\le L \tag{1.8}$$

$$P(x,t)=\int_0^L \rho(x,a,t)da \qquad 0<x<d,\ 0\le t \tag{1.9}$$

$$\rho(x,0,t)=b(x,t)=\int_0^L \beta\rho(x,a,t)da \qquad 0<x<d,\ 0<t \tag{1.10}$$

$$\rho(x,a,0)=\rho(x,a) \qquad 0<x<d,\ 0\le a\le L \tag{1.11}$$

$$\rho(0,a,t)=\rho(d,a,t)=0 \qquad 0\le t,\ 0\le a\le L \tag{1.12}$$

2. IMPLICIT FINITE DIFFERENCE SCHEME

We approximate Eqn.(1.8) by

$$\frac{u_{i,j}^{k+1}-u_{i,j}^{k}}{\Delta t}+\frac{u_{i,j}^{k}-u_{i,j-1}^{k}}{\Delta a}= k\sum_{p=1}^{m}\alpha_p\left[\frac{u_{i-1,p}^{k+1}-2u_{i,p}^{k+1}+u_{i+1,p}^{k+1}}{(\Delta x)^2}\right]-\lambda u_{i,j}^{k+1} \tag{2.1}$$

where

$$u^k_{i,j} \simeq \rho(x_i,a_j,t_k)=\rho(i\Delta x, j\Delta a, k\Delta t),$$

$$i=1,2,\ldots,\ell,\ j=1,2,\ldots,m,\ k=0,1,2,\ldots,$$

and

$$\Delta x=\frac{d}{\ell+1},\quad \Delta a=\frac{L}{m}$$

The sum in (2.1) is a quadrature formula approximating the integral in (1.8). Assuming $\Delta t=\Delta a$, Eqn.(2.1) may be written in the form

$$(1+s)u^{k+1}_{i,j}-r\sum_{p=1}^{m}\alpha_p(u^{k+1}_{i-1,p}-2u^{k+1}_{i,p}+u^{k+1}_{i+1,p})=u^k_{i,j-1} \qquad (2.2)$$

for each $i=1,2,\ldots,\ell$, $j=1,2,\ldots,m$, $k=0,1,\ldots$ where

$$r=\frac{k\Delta t}{(\Delta x)^2},\qquad s=\lambda\Delta t. \qquad (2.3)$$

In view of Eqns.(1.10)-(1.12), the following conditions are imposed on the difference Eqn.(2.2):

$$u^k_{i,0}=b^k_i=b(i\Delta x,j\Delta a)=\beta\sum_{p=1}^{m}\alpha_p u^k_{i,p}, \qquad (2.4)$$

$$u^0_{i,j}=\rho_{i,j}=\rho(i\Delta x,j\Delta a),$$

$$u^{k+1}_{0,j}=u^{k+1}_{\ell+1,j}=0,$$

$$i=1,2,\ldots,\ell,\ j=1,2,\ldots,m,\ k=0,1,2,\ldots\ .$$

The sum in (2.4) is again a quadrature formula approximating the integral in (1.10).

By Eqn.(2.2) we can compute the unknown $u^{k+1}_{i,j}$ in terms of known $u^k_{i,j}$. Collecting the $\ell\times m$ unknowns $u^{k+1}_{i,j}$ and $\ell\times m$ knowns $u^k_{i,j-1}$ in Eqn.(2.2) into column vectors, yields

$$X=\left[u^{k+1}_{1,1}\ldots u^{k+1}_{1,m};\ u^{k+1}_{2,1}\ldots u^{k+1}_{2,m};\ \ldots,\ u^{k+1}_{\ell,1}\ldots u^{k+1}_{\ell,m}\right]^T, \qquad (2.5)$$

$$C=\left[u^k_{1,0}\ldots u^k_{1,m-1};\ u^k_{2,0}\ldots u^k_{2,m-1};\ \ldots,\ u^k_{\ell,0}\ldots u^k_{\ell,m-1}\right]^T, \qquad (2.6)$$

then Eqn.(2.2) is equivalent to a system of linear equations of the form

$$AX=C \qquad (2.7)$$

where A is an $\ell\times m$ by $\ell\times m$ matrix given by

$$A=\begin{bmatrix}(1+s)I+2R & -R & & 0\\ -R & \ddots & \ddots & \\ & \ddots & \ddots & -R\\ 0 & & -R & (1+s)I+2R\end{bmatrix} \qquad (2.8)$$

I is the identity matrix of order m and R is a square matrix of order m which has rank one and is given by

$$R=\begin{bmatrix}r_1 & r_2 & \cdots & r_m\\ r_1 & r_2 & \cdots & r_m\\ \vdots & & & \\ r_1 & r_2 & \cdots & r_m\end{bmatrix} \qquad (2.9)$$

where $r_j=r\alpha_j$, $j=1,2,\ldots,m$.

3. STABILITY

In this section we examin the stability for (2.2) or equivalently for (2.7) by analizing error-growth. Assume at time level k an eror vector E^k was made in representing the vector C. Then an error $A^{-1}E^k$ would de propagated in X; since by Eqn.(2.7), we have

$$X=A^{-1}(C+E^k)=A^{-1}C+A^{-1}E^k.$$

In the next section we will show that A^{-1} exists by showing that none of its eigenvalues is zero.

Let

$$E^{k+1}=A^{-1}E^k, \qquad (3.1)$$

then we have a relation between the errors at time levels k and k+1 and the method defined by Eqn.(2.7) is stable if

$$\|E^{k+1}\|\le\|E^k\|,$$

where $\|\ \|$ is the Euclidian norm.

To find conditions under which (2.7) is stable, in this sense, first we find the eigenvalues and eigenvectors of A in (2.8).

Since R has rank one, N(R), the null space of R, has dimension m-1. Therefore there exists m-1 linearly independent column vectors $w_1,w_2,\ldots,w_{m-1}\in\mathbb{R}^m$ such that $Rw_j=0$, $j=1,2,\ldots,m-1$. Now consider the vectors $v_{i,j}\in\mathbb{R}^{\ell\times m}$ in the form

$$v_{1,j}=\begin{bmatrix}0\\ w_j\\ w_j\\ \vdots\\ w_j\end{bmatrix},\quad v_{2,j}=\begin{bmatrix}w_j\\ 0\\ w_j\\ \vdots\\ w_j\end{bmatrix},\quad \ldots,\quad v_{\ell,j}=\begin{bmatrix}w_j\\ w_j\\ w_j\\ \vdots\\ 0\end{bmatrix},$$

$$j=1,2,\ldots,m-1,$$

where 0 denotes the zero column vector in $\mathbb{R}^m$.

Let us relabel the above $\ell\times(m-1)$ vectors and simply denote them by $v_1,v_2,\ldots,v_{\ell(m-1)}$. These $\ell(m-1)$ vectors are linearly independent and satisfy

$$Av_i=(1+s)v_i,\quad i=1,2,\ldots,\ell(m-1). \qquad (3.2)$$

It follows from Eqn.(3.2) that $\lambda=1+s$ is an eigenvalue of A of multiplicity at least $\ell(m-1)$.

Now we look for ℓ eigenvectors in the form

$$u=\begin{bmatrix} y_1 e \\ y_2 e \\ \vdots \\ y_\ell e \end{bmatrix} \qquad (3.3)$$

where

$$e=\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

and e has m rows. Let λ be the corresponding eigenvalue. Then in $\mathbb{R}^{\ell\times m}$ we have

$$Au=\lambda u. \qquad (3.4)$$

Let $\gamma=\sum_{j=1}^{m} r_j$, then it follows from Eqn.(3.4) that

$$\begin{aligned} (1+s+2\gamma)y_1-\gamma y_2 &= \lambda y_1 \\ -\gamma y_1+(1+s+2\gamma)y_2-\gamma y_3 &= \lambda y_2 \\ &\vdots \\ -\gamma y_{\ell-1}+(1+s+2\gamma)y_\ell &= \lambda y_\ell \end{aligned}$$

In matrix form the above system of ℓ linear equations may be written as

$$BY=\lambda Y, \qquad (3.5)$$

where the matrix B and the vector y have the form

$$B=\begin{bmatrix} 1+s+2\gamma & -\gamma & & \\ -\gamma & 1+s+2\gamma & -\gamma & \\ & \ddots & \ddots & \ddots \\ & & -\gamma & 1+s+2\gamma \end{bmatrix}, \quad y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{bmatrix}$$

Note that B is a square matrix of order ℓ. Eqn.(3.5) shows that λ is also an eigenvalue of B. But the eigenvalues of B are given by

$$\lambda_i=1+s+4\gamma\cos^2\frac{i\pi}{2(\ell+1)}, \quad i=1,2,\ldots,\ell \qquad (3.6)$$

and hence we have obtained the ℓ other eigenvalues of A which (since $\gamma\neq 0$) are obviously distinct.

If u_i $(i=1,2,\ldots,\ell)$ are the eigenvectors corresponding to λ_i $(i=1,2,\ldots,\ell)$, then they are linearly independent. Note that each u_i has the form (3.3). Now we have the sets of eigenvalues and eigenvectors of A which we denote them by T and S respectively

$$T=M\cup N \qquad (3.7)$$

where

$$M=\left\{\lambda_i:\lambda_i=1+s,\ i=1,2,\ldots,\ell(m-1)\right\}$$

$$N=\left\{\lambda_i:\lambda_i=1+s+4\gamma\cos^2\frac{i\pi}{2(\ell+1)},\ i=1,2,\ldots,\ell\right\}$$

and

$$S=\left\{v_1,v_2,\ldots,v_{\ell(m-1)},u_1,u_2,\ldots,u_\ell\right\} \qquad (3.8)$$

Lemma. The set S is linearly independent.

The proof of this lemma will be given in the next section. Let us use the letter v to denote the vectors in S by setting

$$u_1=v_{\ell(m-1)+1},\ u_2=v_{\ell(m-1)+2},\ \ldots,\ u_\ell=v_{\ell(m-1)+\ell},$$

then

$$S=\left\{v_1,v_2,\ldots,v_{\ell m}\right\}.$$

Now we return to the discussion of stability of Eqn.(2.7). The error vector E^k may be expressed uniquely as

$$E^k=\sum_{i=1}^{\ell m} c_i v_i \qquad (3.9)$$

and using Eqn.(3.1), we obtain

$$E^{k+1}=A^{-1}\sum_{i=1}^{\ell m} c_i v_i=\sum_{i=1}^{\ell m} c_i A^{-1}v_i=\sum_{i=1}^{\ell m} c_i\lambda_i^{-1}v_i. \qquad (3.10)$$

From Eqn.(3.9) and (3.10), it follows that the error does not grow if

$$|\lambda_i^{-1}|\leq 1,\ i=1,2,\ldots,\ell m.$$

This condition holds by (3.7) and therefore we have the following result:

Theorem. The implicit formula (2.2) is unconditionally stable.

4. PROOF OF THE LEMMA

Let $C=\left\{v_1,v_2,\ldots,v_{\ell(m-1)}\right\}$ and $D=\left\{u_1,u_2,\ldots u_\ell\right\}$. Let V be the span of C and U be the span of D. Then U and V are subspaces of $\mathbb{R}^{\ell\times m}$ and since v_i $(i=1,2,\ldots,\ell(m-1))$ are linearly independent, we have $\dim V=\ell(m-1)$. Similarly $\dim U=\ell$.

Let

$$W=U+V=\{u+v: u\in U,\ v\in V\},$$

then W also is a subspace of $\mathbb{R}^{\ell\times m}$ and $S\subseteq W$, where S is defined by Eqn.(3.8). We assert that

$$U\cap V=\{0\}, \qquad (4.1)$$

which implies that $W=U\oplus V$, that is, W is the direct sum of U and V.

Let $u^*\in U\cap V$ and $u^*\neq 0$, then u^* can be expressed uniquely as

$$u^* = \sum_{i=1}^{\ell} b_i u_i \qquad (4.2)$$

and

$$u^* = \sum_{i=1}^{\ell(m-1)} c_i v_i \qquad (4.3)$$

where the b_i and c_i are scalers. Multiplying both sides of Eqns.(4.2) and (4.3) by A, we obtain

$$Au^* = \sum_{i=1}^{\ell} b_i Au_i = \sum_{i=1}^{\ell} b_i \lambda_i u_i \qquad (4.4)$$

and

$$Au^* = \sum_{i=1}^{\ell(m-1)} c_i Av_i = \sum_{i=1}^{\ell(m-1)} c_i (1+s)v_i = (1+s)u^* \qquad (4.5)$$

where $\lambda_i (i=1,2,\ldots,\ell)$ are defind by Eqn.(3.7).

From Eqns.(4.4) and (4.5), we can write

$$u^* = \sum_{i=1}^{\ell} \frac{b_i}{1+s} \lambda_i u_i . \qquad (4.6)$$

Since the respresentation of u^* in Eqn.(4.2) is unique, we reach a contradiction. This proves (4.1).

Now we have

$$\dim W = \dim V + \dim U = \ell + \ell(m-1) = \ell m .$$

Hence $W = \mathbb{R}^{\ell \times m}$, and $\ell \times m$ vectors in S which span $\mathbb{R}^{\ell \times m}$ are linearly independent. This completes the proof.

References

[1] H. VON FOERSTER, Some Remarks on Changing Populations: The Kinetics of Cell Proliferation, 382-407, Grune and Stratton, New York, 1959.

[2] M.E. GURTIN and R.C. MacCAMY, Nonlinear age dependent population dynamics, Arch. Rational Mech. Anal. 54(1974), 281-300.

[3] F. HOPPENSTEADT, Mathematical Theories of Populations: Demographics, Genetics and Epidemics, Soc. Ind. Appl. Math., Philadelphia, 1975.

[4] K.E. SWICK, A nonlinear age-dependent model of single species population dynamics, SIMA J. Appl. Math., 32(1977), 484-498.

[5] D.W. ZACHMANN, J.A. LOGAN, Implication of mathematical stability analysis for an Age-Structured Population, submitted to Researches on Population Ecology.

[6] R.C. MacCAMY, A population model with nonlinear diffusion, J. Differential Equations, 39(1981), 52-72.

[7] M.E. GURTIN, Some questions and open problems in continuum mechanics and population dynamics, J. Differential Equations, 48(1983), 293-312.

[8] A. KERAYECHIAN, D.W. ZACHMANN, Existence of solutions to models of age-dependent pulations with finite life span, J. Math. Anal. and Appl., 2(1986), 403-421.

AUTHOR INDEX