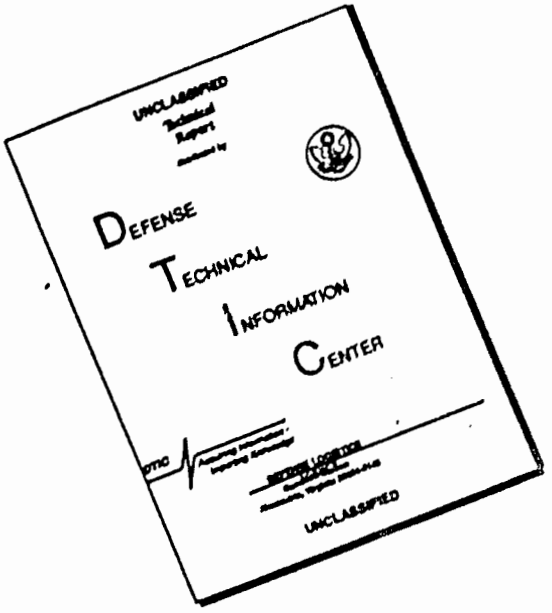
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 17, NUMBER 3

Office of Naval Research
Arlington, Virginia

September 1970

DISCLAIMER NOTICE



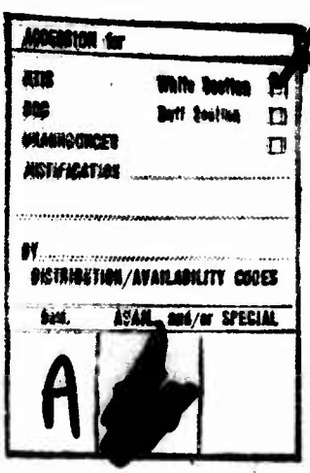
THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NAVAL RESEARCH LOGISTICS QUARTERLY

SEPTEMBER 1970
VOL. 17, NO. 3

OFFICE OF NAVAL RESEARCH

NAVSO P-1278



NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

OPTIMAL INTERDICTION OF A SUPPLY NETWORK

Alan W. McMasters

and

Thomas M. Mustin, LCdr., USN

Naval Postgraduate School
Monterey, California

ABSTRACT

Under certain conditions, the re-supply capability of a combatant force may be limited by the characteristics of the transportation network over which supplies must flow. Interdiction by an opposing force may be used to reduce the capacity of that network. The effects of such efforts vary for differing missions and targets. With only a limited total budget available, the interdictor must decide which targets to hit, and with how much effort. An algorithm is presented for determining the optimum interdiction plan for minimizing network flow capacity when the minimum capacity on an arc is positive and the cost of interdiction is a linear function of arc capacity reduction.

The problem of reducing the maximum flow in a network has received considerable interest recently [1, 3, 8, 9], primarily as a consequence of the problem of interdicting supply lines in limited warfare. In this paper an algorithm is presented for reducing the maximum flow in such a network when the resources of the interdicting force are limited. A typical problem is that of the strike planner who must determine the best way to allocate a limited number of aircraft to interdict an enemy's supply lines on a particular day.

The network is assumed to be capacity limited and to be representable as a planar connected graph of nodes and undirected capacitated arcs. Further, it is assumed to have a single source through which flow enters the network and a single sink through which flow leaves. The maximum flow through such networks is easily determined by finding the minimum cut set where a cut set is defined as a set of arcs which, when removed, causes a network to be partitioned into two subgraphs, one subgraph containing the source node and the other containing the sink node. The value of a cut set is the sum of the flow capacities of its arcs. The minimum cut set is that cut set whose value is the minimum of all cut sets of a network. The max-flow min-cut theorem states that the maximum flow possible through the network is equal to the value of the minimum cut set [4, 5].

In the interdiction problem, an arc (i, j) is assumed to have a maximum flow capacity, $u_{ij} \geq 0$, and a minimum flow capacity, $l_{ij} \geq 0$. At least one arc of the network is assumed to have $l_{ij} > 0$. As a consequence of interdiction, the actual capacity, m_{ij}, on an arc will be somewhere in the range $0 \leq l_{ij} \leq m_{ij} \leq u_{ij}$.

If we assume that the interdictor incurs a cost, C_{ij}, per unit of capacity decrease, then his total cost for reducing an arc's capacity from u_{ij} to m_{ij} will be $C_{ij}[u_{ij} - m_{ij}]$. If we assume the interdictor has a total budget limitation, K, which he cannot exceed, then

$$\sum_{\text{all } (i,j)} C_{ij}[u_{ij} - m_{ij}] \leq K.$$

The cost, C_{ij}, might represent the number of sorties required to reduce arc capacity by one unit and K might represent the total number of sorties which can be flown in a 24-hour period.

The interdictor's problem is to find a set of m_{ij} which minimizes the maximum flow in the supply network subject to

$$\sum_{\text{all } (i,j)} C_{ij}[u_{ij} - m_{ij}] \leq K$$

and

$$l_{ij} \leq m_{ij} \leq u_{ij} \qquad \text{for all } (i,j).$$

Topological Dual

In resolving the interdictor's problem we will make use of the topological dual. This dual, when defined, is another network in which the arcs have lengths instead of capacities. A one-to-one correspondence exists between the cut sets of the original or primal network and the loopless paths through the dual. The problem of finding the minimum cut set in the primal is equivalent to finding the shortest path through the dual [4].

Let the original maximum flow network be called the primal. To construct the topological dual we begin by adding an artificial arc connecting the source to the sink in the primal. The resulting network will be referred to as the modified primal and the area surrounding this network will be referred to as the external mesh. A dual is defined if and only if the modified primal is planar; a planar network being one that can be drawn on a plane such that no two arcs intersect except at a node.

When defined, a dual may be constructed for the interdiction problem in the following manner [9]:

1. Place a node in each mesh of the modified primal including the external mesh. Let the source of the dual be the node in the mesh involving the artificial arc and the sink be the node in the external mesh.

2. For each arc in the primal (except the artificial arc) construct an arc that intersects it and joins with nodes in the meshes adjacent to it.

3. Assign each arc of the dual a length equal to the capacity of the primal arc it intersects.

Preview of the Algorithm

The algorithm begins by ignoring the budget restriction. All arcs of the primal are initially assigned capacities l_{ij} and the shortest route through the topological dual is determined. The length of the route corresponds to the value of the minimum cut set of the primal when $m_{ij} = l_{ij}$ for all arcs. A check is then made to determine if the interdiction cost for obtaining this minimum cut exceeds the budget constraint. If not, then the problem is solved. If, however, the budget constraint has been exceeded then a reduction in expenditures is required.

The algorithm seeks to "unspend" as carefully as possible so that the amount of flow through the network increases as little as possible. The first step in this unspending operation is to find which arc of the minimum cut set "gives back" the largest amount of expense for the smallest increase in capacity. Unspending takes place until $m_{ij} = u_{ij}$ or the budget constraint is satisfied. If $m_{ij} = u_{ij}$ then the algorithm continues working on the minimum cut set until the budget constraint is satisfied. The final value of that cut set is then determined and retained for later comparisons.

The algorithm looks next for the second shortest route corresponding to the second lowest valued cut set when all arcs have $m_{ij} = l_{ij}$. It repeats the budget check and the unspending process. After the budget is satisfied on this cut set then the cut set value is compared with the final value of the cut set of the "shortest" routes; that cut set having the lower final value is retained and the other is dropped from further consideration.

The process continues with consideration next of the third shortest route or third minimum cut set with all arcs having $m_{ij}=l_{ij}$ and then the fourth and so on. If, at any time, the length of the next shortest route using all l_{ij}'s is greater than the final length of the best previous route, the algorithm terminates. There is no point in continuing the next shortest route investigations since all further routes will have lengths greater than the feasible length of the best previous route.

Feasible Min-Cut Algorithm

1. Construct the topological dual of the network and set all $m_{ij}=l_{ij}$. Set $r=1$.

2. Determine R_r, the rth shortest loopless route through the dual when $m_{ij}=l_{ij}$, and determine its length L_r^* from

$$L_r^*=\sum_{(i,j)\in R_r} l_{ij}.$$

If $w \geq 2$ routes qualify for the rth shortest route because of ties in total length, arbitrarily select one of these routes as the rth, another as the $(r+1)$th, another as the $(r+2)$th, and so on, with the last of the group being designated as the $(r+w-1)$th shortest route.

Compare L_r^* with $L^{(r-1)}$, the length of the shortest feasible route from the set $R_1, R_2, \ldots, R_{r-1}$. (Let $L^{(0)}=\infty$).

(a) If $L_r^* < L^{(r-1)}$ then go to step 3.

(b) If $L_r^* \geq L^{(r-1)}$ then terminate the algorithm. The routes $R_r, R_{r+1}, R_{r+2}, \ldots, R_N$ will have feasible lengths which are no shorter than $L^{(r-1)}$ and need not be considered.

3. Compute the interdiction expense, E_r, associated with L_r^* from

$$E_r=\sum_{(i,j)\in R_r} C_{ij}[u_{ij}-l_{ij}].$$

(a) If $E_r \leq K$, terminate the algorithm. Route R_r has the minimum feasible length of all routes through the dual.

(b) If $E_r > K$, go to step 4.

4. List the n arcs in R_r in descending order of C_{ij} values; let $C_1(r)$ represent the largest C_{ij} and $C_n(r)$, the lowest. Beginning with $q=1$ and $L_r=L_r^*$, increase the length of the arc (i, j) corresponding to $C_q(r)$ and the route length L_r by

$$\Delta m_{ij}=\min\left\{u_{ij}-l_{ij}, \frac{E_r-K}{C_{ij}}, L^{(r-1)}-L_r\right\}.$$

Decrease the interdiction expense E_r by $C_{ij}\Delta m_{ij}$.

(a) If $\Delta m_{ij}=u_{ij}-l_{ij}$ increase q by 1; compute Δm_{ij} and the new values of L_r and E_r for the next arc on the C_{ij} list.

(b) If $\Delta m_{ij}=\dfrac{E_r-K}{C_{ij}}$, the interdiction expense for the route is $E_r=K$. If $L_r \leq L^{(r-1)}$, set $L^{(r)}=L_r$ and record the current value of q, call it s. Delete the route associated with $L^{(r-1)}$ from further consideration. If $L_r > L^{(r-1)}$, set $L^{(r)}=L^{(r-1)}$ and drop R_r from further consideration. Increase r by 1 and return to step 2.

(c) If $\Delta m_{ij}=L^{(r-1)}-L_r$, the length of route r has been increased to $L^{(r-1)}$, but it is still not feasible since $E > K$. Delete R_r from further consideration, set $L^{(r)}=L^{(r-1)}$, and return to step 2.

If there is a tie between $u_{ij}-l_{ij}$ or $L^{(r-1)}-L_r$ and $\frac{E_r-K}{C_{ij}}$ for value of Δm_{ij}, apply part (b) above. If there is a tie between $u_{ij}-l_{ij}$ and $L^{(r-1)}-L_r$, apply part (c).

Optimal Allocation

The value of $L^{(r)}$ at the termination of the algorithm is the minimum value of all the feasible cut sets. This is the minimum achievable network capacity. The interdiction effort is assigned to the arcs of the primal which are "cut" by the feasible route R_p of the topological dual associated with the value of $L^{(r)}$. The optimal number of sorties to allocate is

$$n_{ij}=C_{ij}[u_{ij}-l_{ij}]$$

for the arcs of the primal cut by the dual arcs of R_p associated with $C_{s+1}(p)$, $C_{s+2}(p)$, . . . , $C_n(p)$ where s is the index from the C_{ij} list of the first arc on R_p having $\Delta m_{ij}>0$. For the arc (i,j) associated with $C_s(p)$:

$$n_{ij}=K-\sum_{C_{s+1}(p)}^{C_n(p)} n_{ij}.$$

Finally, $n_{ij}=0$ for all other arcs of the primal network.

EXAMPLE: Figure 1 presents the network information for the example. The value of K will be 5. Node 1 is the source and node 5 is the sink. The numbers on each arc represent l_{ij}, u_{ij}; C_{ij}.

The topological dual is formed as shown by the dashed lines in Figure 1. The artifical arc added to the primal for constructing the dual is arc (5, 1). The completed topological dual is shown in Figure 2; the numbers on the arcs represent the upper and lower bounds on arc length and the unit costs for shortening them. These numbers correspond directly to the numbers on the arcs of the primal cut by the dual arcs. The source and sink of the dual are nodes A and D, respectively.

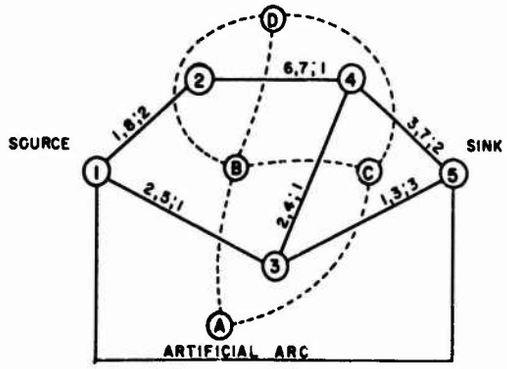


FIGURE 1. A supply network

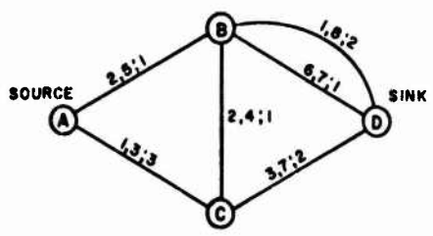


FIGURE 2. The topological dual of the network

When $m_{ij} = l_{ij}$ on all of the arcs of the dual the complete set of loopless routes from source to sink with associated lengths L_r^* can be obtained by inspection. It is:

R_1: (AB, BD1) $L_1^* = 3$
R_2: (AC, CD) $L_2^* = 4$
R_3: (AC, CB, BD1) $L_3^* = 4$
R_4: (AB, BC, CD) $L_4^* = 7$
R_5: (AB, BD2) $L_5^* = 8$
R_6: (AC, CB, BD2) $L_6^* = 9.$

The designation BD1 is associated with the upper BD arc in Figure 2 and BD2 is associated with lower. Although the algorithm would not evaluate all routes R_1 through R_6 and their associated L_r^* values they are presented for the sake of discussion.

The algorithm begins by finding R_1 and computing $L_1^* = 3$. $L^{(0)} = \infty$ is set so that $L_1^* < L^{(0)}$. Because $E_1 = 17 > K$, the cost coefficients for R_1 are ranked, $C_1(1) = 2$ (for arc BD1) and $C_2(1) = 1$ (for arc AB). The evaluation of Δm_{BD1} results in

$$\Delta m_{BD1} = \frac{E_1 - K}{C_{BD1}} = 6,$$

$L_1 = 9$, and $E_1 = 5 = K$. The analysis of R_1 is complete because $E_1 = K$, therefore $L^{(1)} = L_1 = 9$.

After finding R_2, the value L_2^* is computed. Because $L_2^* = 4 < L^{(1)}$, the value of E_2 is next determined. $E_2 = 14 > K$ so the cost coefficients for R_2 must be ranked. $C_1(2) = 3$ (for arc AC) and $\Delta m_{AC} = u_{AC} - l_{AC} = 2$ resulting in $L_2 = 6$ and $E_2 = 8$. Next $\Delta m_{CD} = \dfrac{E_2 - K}{C_{CD}} = 3/2$ so $L_2 = 7\frac{1}{2}$ and $E_2 = 5 = K$, completing the analysis of R_2.

Because $L_2 < L^{(1)}$ we drop R_1 from further consideration and set $L^{(2)} = L_2 = 7\frac{1}{2}$.

R_3 is next on the list. $L_3^* < L^{(2)}$ so E_3 is determined. $E_3 = 22 > K$ and Δm_{AC} must then be calculated. We get $\Delta m_{AC} = u_{AC} - l_{AC} = 2$ resulting in $L_3 = 6$ and $E_3 = 16$. Next, $\Delta m_{BD1} = L^{(2)} - L_3 = 3/2$ and R_3 can be disregarded. Set $L^{(3)} = L^{(2)} = 7\frac{1}{2}$.

Route R_4 has $L_4^* = 7 < L^{(3)}$ and $E_4 = 13$. Then $\Delta m_{CD} = L^{(3)} - L_4 = \frac{1}{2}$ and we can disregard R_4. Set $L^{(4)} = L^{(3)} = 7\frac{1}{2}$.

Because R_5 has $L_5^* = 8 > L^{(4)}$ the algorithm terminates.

The dual route which is used to determine the optimal allocation of interdiction effort is R_2. $L_2 = 7\frac{1}{2}$ is the value of the minimum cut of the primal network after optimal interdiction. Arc AC has length $m_{AC} = u_{AC} = 3$ and arc CD has a length $m_{CD} = 4\frac{1}{2} < u_{CD}$. Therefore arc (3, 5) of the primal has a final capacity of $m_{35} = u_{35} = 3$ and arc (4, 5) of the primal has a final capacity of $m_{45} = 4\frac{1}{2}$. The entire budget $K = 5$ is allocated to interdiction of arc (4, 5). This optimal interdiction gives a maximum possible flow through the network of $7\frac{1}{2}$.

An *r*th Shortest Route Algorithm

An algorithm for finding the *r*th shortest loopless route through the dual network is a necessary part of step 2 of the Feasible Min-Cut algorithm for large problems. Such an algorithm can be derived by minor modifications to the "N best loopless paths" algorithm of Clarke, Krikorian, and Rausen [2] (their algorithm will be referred to as the CKR algorithm from this point on). In seeking the N best loopless paths the CKR algorithm concentrates on paths which have at most one loop. The procedure

begins with the determination of an initial set S of N loopless routes along with a set T of routes having one loop, but lengths less than the longest of the N routes of S. Special deviations, called "detours," from routes in the set T are then examined to see if any loopless route arises which is shorter in length than the longest of set S. If so, then this route replaces the longer one in S. When the elements of sets S and T cease changing the algorithm terminates.

The modification for converting this procedure to an rth shortest route type is quite simple. Use the CKR algorithm to find an initial set of $N \geq 1$ best loopless routes. If, during the course of applying the Feasible Min-Cut algorithm additional routes beyond N are needed, use the existing N routes to initiate the construction of the new set S. The new set S is initially established when a specified number of loopless routes, $K(\geq 1)$, has been added to S. Those detours of routes in new S having loops, but total lengths less than the maximum from S form the new set T. The CKR algorithm is then applied to find the final set of $N+K$ best loopless routes.

If more than $N+K$ routes are needed after returning to the Feasible Min-Cut algorithm then another set of K additional routes can be added in the same way as the first K. The second new set S would be initiated with the existing $N+K$ best loopless routes.

The values of N and K are a matter of personal choice. The use of $K=1$ does not however seem very efficient because of the possibility of multiple routes of the same length. With $K>1$ such ties become more quickly apparent. In any case, a complete list of all routes of a particular length should be evaluated before returning to the Feasible Min-Cut algorithm. For example, if there are three shortest routes through the network and $N=2$ was used then an additional set of $K \geq 2$ routes should be evaluated to pick up the third route and to show that there is only one more shortest route prior to going to step 3 of the Feasible Min-Cut algorithm.

Modifications when all $l_{ij}=0$

The Feasible Min-Cut algorithm was designed for problems where at least one arc has $l_{ij}>0$. The reason for this was that in most real-world interdiction problems it would be virtually impossible to reduce an arc's capacity to zero for any extended period of time [3, 6]. Often hand-carrying of supplies can begin immediately after an aerial or ground attack. If one considers l_{ij} to represent the average 24 hour minimum capacity then hand-carrying and minor repairs would definitely result in $l_{ij}>0$.

If the Feasible Min-Cut algorithm is applied to a network having all $l_{ij}=0$ it would evaluate the feasible length of all loopless routes through the dual. The following modifications in steps 1 and 2 of the algorithm are suggested as a means of possibly avoiding this complete evaluation. Step 3 would be by-passed completely.

1. Construct the topological dual of the network and set all $m_{ij}=u_{ij}$. Set $r=1$.

2. Determine R_r, the rth shortest loopless route through the dual when $m_{ij}=u_{ij}$. Then set $m_{ij}=0$ for all arcs on this route and determine E_r from

$$E_r=\sum_{(i,j)\in R_r} C_{ij}u_{ij}.$$

(a) If $E_r \leq K$, terminate the algorithm. Route R_r has a minimum feasible length of zero and $n_{ij}=C_{ij}u_{ij}$ for all arcs on R_r.

(b) If $E_r > K$, go to step 4.

Comments

The algorithm terminates in a finite number of steps since the number of loopless routes through the dual network is finite for finite networks and each route is examined only once.

If all l_{ij}, u_{ij}, C_{ij}, as well as K are integer valued then n_{ij} will be integer also. If any of these parameters is not integer then there is no guarantee of an integer solution. If a problem involves allocating sorties then integer solutions should be sought after the Feasible Min-Cut algorithm is completed. If, however, the problem involves allocating, say, tons of bombs, then noninteger results might be quite reasonable.

Extensions

The law of diminishing returns suggests that actual interdiction costs for an arc (i, j) may follow a curve of the type shown in figure 3. The Feasible Min-Cut algorithm can solve problems having this type of nonlinear cost function if the function is replaced by a piecewise linear approximation such as that shown by the dashed lines in Figure 3. This linear approximation can be created in the primal network by replacing arc (i, j) by three arcs having l_{ij}, u_{ij}, and C_{ij} values as shown in Figure 4. The construction of the topological dual will then require that a node be placed in each mesh of Figure 4.

A further extension of the interdiction problem with nonlinear costs has been made by Nugent [7]. He considers an exponential cost function in continuous form and presents an algorithm similar to the Feasible Min-Cut algorithm for solving the problem.

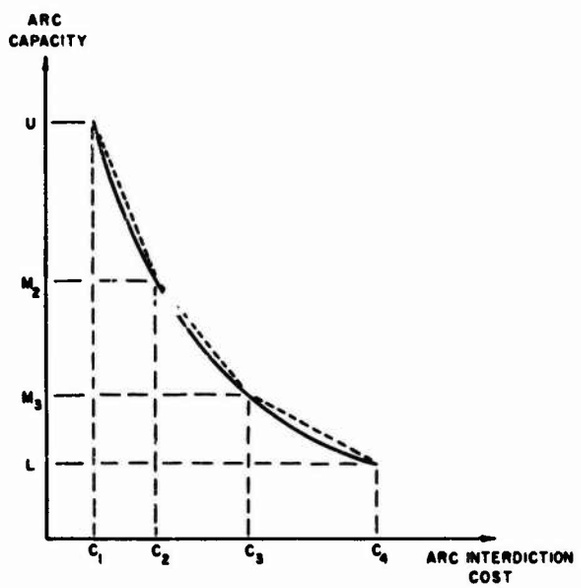


FIGURE 3. Arc capacity as a function of interdiction cost under the law of diminishing returns

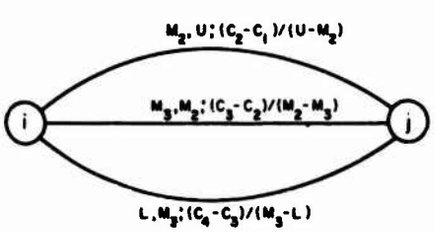


FIGURE 4. Replacement of arc (i, j) for the linear approximation to Fig. 3

REFERENCES

[1] Bellmore, M., J. J. Greenberg, and J. J. Jarvis, "Optimal Attack of a Communications Network," Paper WA 2.4, presented at the 32d National ORSA Meeting, Chicago, November 1967.

[2] Clarke, S., A. Krikorian, and J. Rausen, "Computing the N Best Loopless Paths in a Network," J. SIAM **11**, 1096–1102 (1963).

[3] Durbin, E. P., "An Interdiction Model of Highway Transportation," The RAND Corporation, Rpt. RM–4945–PR (1966).

[4] Ford, L. R. and D. R. Fulkerson, "Maximal Flow through a Network," Canadian J. Math. **8**, 399–404, 1956

[5] Ford, L. R. and D. R. Fulkerson, *Flows in Network* (Princeton Univ. Press, Princeton, N.J., 1962).

[6] Futrell, R. F., *The United States Air Force in Korea, 1950–1953* (Duell, Sloan, and Pearce, New York, 1961).

[7] Nugent, R. O., "Optimum Allocation of Air Strikes Against a Transportation Network for an Exponential Damage Function," Unpublished Masters' Thesis, Naval Postgradaate School, 1969.

[8] Thomas, C. J., "Simple Models Useful in Allocating Aerial Interdiction Effort," Paper WP4.1, 34th National ORSA Meeting, Philadelphia, Nov. 1968.

[9] Wollmer, R. D., "Removing Arcs From a Network," Operations Research **12**, 934–940 (1964).

OPTIMAL MULTICOMMODITY NETWORK FLOWS WITH RESOURCE ALLOCATION

J. E. Cremeans, R. A. Smith and G. R. Tyndall

Research Analysis Corporation
McLean, Virginia

ABSTRACT

The problem of determining multicommodity flows over a capacitated network subject to resource constraints may be solved by linear programming; however, the number of potential vectors in most applications is such that the standard arc-chain formulation becomes impractical. This paper describes an approach—an extension of the column generation technique used in the multicommodity network flow problem—that simultaneously considers network chain selection and resource allocation, thus making the problem both manageable and optimal. The flow attained is constrained by resource availability and network capacity. A minimum-cost formulation is described and an extension to permit the substitution of resources is developed. Computational experience with the model is discussed.

INTRODUCTION

The problem of multicommodity flows in capacitated networks has received considerable attention. Ford and Fulkerson [3] suggested a computational procedure to solve the general maximum-flow case. Tomlin [7] has extended the procedure to include the minimum-cost case. Jewell [6] has pointed out the strong historical and logical connection between this solution procedure for the multicommodity problem and the decomposition algorithm of Dantzig and Wolfe [2].

A related problem, which has not been directly addressed, is the determination of multicommodity flows in a system constrained by resource availability. For example, flows in transportation networks are constrained by available resources that must be shared by two or more arcs in the network. The determination of the set of routes and the allocation of resources to these routes to maximize multicommodity flows or to minimize system cost in meeting fixed flow requirements can be applied to many problems in logistics and other areas. This paper discusses a solution procedure for multicommodity network flows with resource constraints in a minimum-cost case and develops an extension to permit the substitution of resources.

THE MULTICOMMODITY NETWORK FLOW PROBLEM

Consider the multimode, multicommodity network $G(N, \mathscr{A})$. N is the set of all the nodes of the network. $\mathscr{A}$ is the subset of all ordered pairs (x, y) of the elements of N that are arcs of the network. $\mathscr{A}_1, \ldots, \mathscr{A}_m$ is an enumeration of the arcs. Each arc has an associated capacity $b_{(x,y)} \geq 0$ and an associated cost (or distance) $d_{(x,y)} \geq 0$.

For each commodity $k (k = 1, \ldots, q)$ there is a source s_k and a sink i_k. The flow of commodity k along a directed arc (x, y) is $F^k_{(x,y)}$, $(k = 1, \ldots, q)$, and these $F^k_{(x,y)}$, $(k = 1, \ldots, q)$ must satisfy the capacity constraints

$$\sum_{k=1}^{q} F^k_{(x,y)} \leq b_{(x,y)} \quad [(x, y) \in \mathscr{A}].$$

The multicommodity network flow problem as formulated by Ford and Fulkerson [3] is as follows: Define the set $P^k=\{P_j^{(k)}|P_j^{(k)}$ is a chain connecting s_k and $t_k\}$. Now let P be the union of the sets P^k $(k=1, \ldots, q)$. Further, let $P_1^{(1)}, P_2^{(1)}, \ldots, P_j^{(k)}, \ldots, P_n^{(q)}$ be the enumeration of the chains $P_j^{(k)}\epsilon P$ such that the subscript j is sufficient to identify the chain, its origin-destination pair, and the commodity with which it is associated.

Thus the kth commodity set is defined by

$$J_k=\{j|P_j^{(k)} \text{ is a chain from } s_k \text{ to } t_k\},\ k=1, \ldots, q.$$

The arc-chain incidence matrix is

$$A=[a_{ij}],$$

where

$$a_{ij}=\begin{cases}1 \text{ if } \mathscr{A}_i\epsilon P_j^{(k)}\\ 0 \text{ otherwise}\end{cases}$$

for $i=1 \ldots, m$; $j=1, \ldots, n$. Each column of the matrix A is thus a representation of a chain $P_j^{(k)}$.

Consider the network used as an example in Ref [3], augmented by s_k and t_k $(k=1, 2)$, with source s_1 and sink t_1 for commodity 1 and source s_2 and sink t_2 for commodity 2. Figure 1 illustrates the network and Figure 2 shows the arc-chain incidence matrix A.

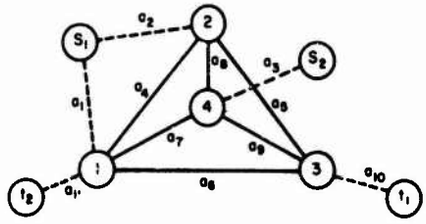


FIGURE 1. Network A

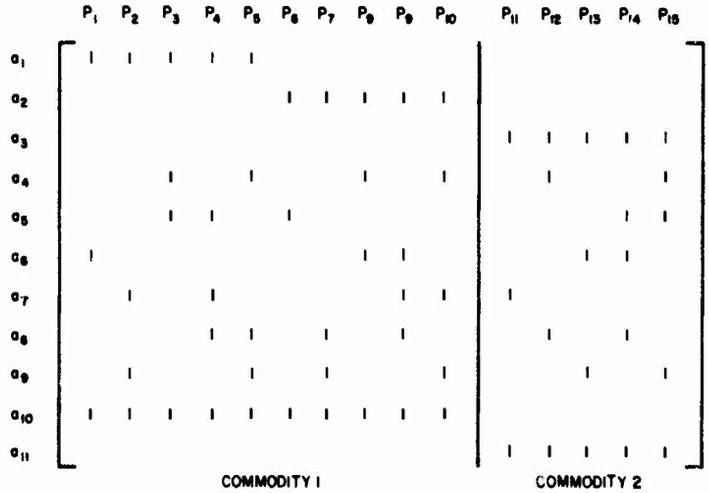


FIGURE 2. Arc-Chain Incidence Matrix A

Letting $x_j^{(k)}$ $(j=1, \ldots, n)$ be the flow of commodity k in chain $P_j^{(k)}$ $(j=1, \ldots, n;\ k$ implicit) and b_i the flow capacity of $\mathscr{A}_i$, the multicommodity, maximum-flow linear program is:
Maximize

$$\sum_{j=1}^{n} x_j^{(k)}$$

subject to capacity constraints

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i \text{ for } i=1, \ldots, m.$$

Thus the objective is to maximize flow over all possible chains from origins to their respective destinations subject to the capacity constraints of the arcs.

The number of variables in the aforementioned linear program is very large since the number of possible chains is very large in most applications. The procedure proposed by Ford and Fulkerson [3] is to treat the nonbasic variables implicitly; i.e., nonbasic chains are not enumerated. The column vector to enter the basis is generated by applying the simplex multipliers to the arcs as pseudo costs and selecting the candidate chain using the shortest chain algorithm.*

Extension To Include Resource Constraints

In the linear programming problem stated previously, flow is to be maximized subject to the constraints imposed by the capacities of the individual arcs of the network. In some applications additional constraints on flow are imposed by the limited availability of resources used jointly by two or more arcs of the network. An example of this type of network, which will be used throughout the remainder of this paper, is a transportation network.

It is clear that the simultaneous consideration of both types of constraints is an important problem in transportation networks. Roadways, rail lines, etc., have capacity limitations that may limit the maximum movement of men and materials, particularly in less-developed areas. The vehicles and resources available to use the network can actually impose a greater constraint on total movement than the arc capacities. In a highly developed transportation system the capacity of the network may greatly exceed that required; the effective limitations of movement result from too few vehicles or other resources.

For the purposes of this paper, resources are defined to be men, equipment, or other mobile assets that are required to accomplish flow on many arcs of the network. For example, trucks, locomotives, labor, etc., are resources in a transportation network. To effect the simultaneous consideration of resource and network capacities, we may represent resource requirements as follows:

Let the resource matrix for commodity k be

$$R^k=[r_{is}^k]\,(i=1, \ldots, m;\ s=1, \ldots, p),$$

where r_{is}^k is the quantity of resource s required to sustain a unit flow of commodity k over arc i; $r_{is}^k \geq 0$. Note that for some arc commodity combinations

$$r_{is}^k=\infty (s=1, \ldots, p),$$

e.g., if the arc represents a pipeline and the commodity is passengers.

*Professor Mandell Bellmore of The Johns Hopkins University and Mr. Donald Boyer, formerly of the Logistics Research Project, The George Washington University, have developed computer programs to solve the problem using this procedure.

Letting ρ_s be the quantity of resource s available (e.g., in inventory) for assignment to the network ($s=1, \ldots, p$), the minimum-cost multi-commodity network flow problem with resource constraints may be formulated in arc-chain terms as follows:

Minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)}$$

subject to

(a) capacity constraints

$$\sum_{j=1}^{n} a_{ij} x_j^{(k)} \leq b_i \text{ for } i=1, \ldots, m$$

(b) resource constraints

$$\sum_{i=1}^{m} \sum_{k=1}^{q} \sum_{j \in J_k} a_{ij} x_j^{(k)} r_{is}^k \leq \rho_s \text{ for } s=1, \ldots, p$$

and

(c) delivery requirements

$$\sum_{j \in J_k} x_j^{(k)} = \lambda_k \text{ for } k=1, \ldots, q$$

where λ_k is the delivery requirement at t_k ($k=1, \ldots, q$), ($\lambda_k \geq 0$).*

The cost coefficient, c_j, may be defined as:

$$c_j = \sum_{i=1}^{m} \tau_i a_{ij} + \sum_{s=1}^{p} \sum_{i=1}^{m} \phi_s r_{is}^k a_{ij} \text{ (for } j=1, \ldots, n; k \text{ where } P_j^{(k)} \text{ connects } s_k \text{ and } t_k),$$

where τ_i is the cost (or toll) for a unit flow over arc i, and ϕ_s is the cost of using a unit of resource s.

Define the matrices G and $\hat{A}$ as follows: G is a commodity delivery incidence matrix ($q \times n$)

$$G = [g_{kj}]$$

where

$$g_{kj} = \begin{cases} 1 \text{ if } j \in J_k \\ 0 \text{ otherwise} \end{cases}.$$

$\hat{A}$ is a matrix ($m+p+q \times n$) formed of the submatrices A, E, and G as follows:

$$\hat{A} = \begin{bmatrix} A \\ \hline E \\ \hline G \end{bmatrix}$$

The typical column of $\hat{A}$ is

$$\hat{A}_j = \text{col. } (a_{1j}, \ldots, a_{mj}, e_{1j}, \ldots, e_{pj}, g_{1j}, \ldots, g_{qj}).$$

*The case where $p=1$ is equivalent to the "arc-chain formation" of Ref [7].

Solution Procedure Using the Column Generation Technique

The minimum-cost linear program with arc capacity and resource constraints and delivery requirements will be quite large for most applications and will, in addition, require considerable preliminary computation to obtain the coefficients of the $\hat{A}$ matrix. (The authors have solved several small problems using a standard linear programming code.) The column generation procedure suggested by Ford and Fulkerson [3] can be modified to apply to the problem extended to include resource constraints and delivery requirements so that it is never necessary to form the $\hat{A}$ matrix explicitly. The shortest chain algorithm [4] can be used to develop the $\hat{A}_j$ that will satisfy the simplex rule. Further, if the shortest chain algorithm can find no chain satisfying the requirement, an optimum has been reached.

This formulation can be solved by adopting the standard two-phased procedure. Phase I minimizes to zero the value of

$$\sum_{j=n+m+p+1}^{j=n+m+p+q} x_j^{(k)},$$

to obtain an initial basic feasible solution. This effectively assigns a cost of 1 to the artificial variables and a cost of zero to the other variables in Phase I. Phase II begins with the basic feasible solution determined in Phase I and proceeds to minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)}.$$

In Phase I, I_{m+p+q} may be used as the initial basis and the simplex rule is to enter a chain in the basis if, and only if,

$$c_j - c_B B^{-1} \hat{A}_j < 0,$$

where

$$c_B B^{-1} = (\alpha_1, \ldots, \alpha_m, \pi_1, \ldots, \pi_p, \sigma_1, \ldots, \sigma_q),$$

so that the simplex multipliers α_i are associated with the arcs, the π_s are associated with the resources, and the σ_k are associated with the artificial variables. Thus the vector $\hat{A}_j$ is entered if

$$-\sum_{i=1}^{m} a_{ij}\left[\alpha_i + \sum_{s=1}^{p} \pi_s r_{is}^k\right] < \sigma_k.$$

Thus the contribution of each arc to $c_j - c_B B^{-1} \hat{A}_j$ in Phase I is

$$d_i^k = -\left[\alpha_i + \sum_{s=1}^{p} \pi_s r_{is}^k\right].$$

We may use the shortest chain algorithm to find

$$\min_j \left[\sum_{\mathscr{A}_i \in P_j} d_i^k\right] = \min_j \left[\sum_{\mathscr{A}_i \in P_j} \left(-\alpha_i - \sum_{s=1}^{p} \pi_s r_{is}^k\right)\right] \text{ over all } k.$$

Where $\mathcal{A}_i$ is the ith arc and P_j is the jth chain from s_k to t_k. The minimum over all commodities is selected as the candidate to enter the basis. The column vector to leave the basis may be determined in the standard simplex fashion. Should any α_i or π_s be positive the corresponding slack variable could be entered into the basis.

In Phase II

$$c_j - c_B B^{-1} \hat{A}_j = \sum_{i=1}^{m} \tau_i a_{ij} + \sum_{s=1}^{p} \phi_s e_{sj} - \sum_{i=1}^{m} \alpha_i a_{ij} - \sum_{s=1}^{p} \pi_s e_{sj} - \sum_{k=1}^{q} \sigma_k g_{kj}$$

$$= \sum_{i=1}^{m} a_{ij}(\tau_i - \alpha_i) + \sum_{s=1}^{p} \sum_{i=1}^{m} a_{ij} r_{is}^k (\phi_s - \pi_s) - \sum_{k=1}^{q} \sigma_k g_{kj},$$

where

$$e_{sj} = \sum_{i=1}^{m} r_{is}^k a_{ij},$$

and

$$g_{kj} = \begin{cases} 1 \text{ if } j \epsilon J_k \\ 0 \text{ otherwise} \end{cases}.$$

Thus

$$d_i^k = \tau_i - \alpha_i + \sum_{s=1}^{p} r_{is}^k (\phi_s - \pi_s)$$

may be assigned to arc i. The shortest, i.e., the chain with the least,

$$\sum_{i=1}^{m} d_i^k - \sigma_k < 0$$

for $k = 1, \ldots, q$, may then be entered in the basis.

Phase II is terminated and the value of z is minimized when, for the minimum j,

$$\sum_{\mathcal{A}_i \epsilon P_j} d_i^k \geq \sigma_k.$$

Extension for Substitution of Resources

In the previous section only one combination of resources was permitted to be applied to an arc in order to move one unit of commodity k over arc i. We now present a modification of the initial formulation to allow for the substitution of resources. In economic terms the arc-commodity pair is similar to a production function with constant returns to scale and fixed technical coefficients (see Ref. [1], p. 36). In some applications this may be a significant limitation. Consider again a transportation network. A highway arc might be considered for the transport of manufactured products. Closed vans with a driver and an alternate driver might be the most efficient combination of resources. A combination of a van and one driver would be less efficient, perhaps, since more rest periods would be required, but it is nevertheless a feasible combination. Similarly a third alternative would be the utilization of stake and platform trucks with containers, possibly more expensive than the first two alternatives, but still feasible.

Specific inventories of trucks and drivers may exist and the objective might be to assign these sets of resources in the most efficient way over all arcs even if some arcs or commodities are assigned a less than most efficient set of resources.

In the previous formulation, for each arc-commodity pair a single combination of resources is required and represented by the vector, $R_i^k = (r_{i1}^k, r_{i2}^k, \ldots, r_{ip}^k)$, of the matrix R^k.

Define a new resource matrix: $T = [t_{ik}] (i = 1, \ldots, m; k = 1, \ldots, q)$, where $t_{ik} = \{\hat{R}_i^k | \hat{R}_i^k$ is any feasible resource vector for arc i, commodity $k\}$; $\hat{R}_i^k = (\hat{r}_{i1}^k, \ldots, \hat{r}_{ip}^k)$.

In words, each element of T is the set of alternative resource vectors for a movement of one unit of commodity k over arc i.

The contribution of each arc to $c_j - c_B B^{-1} \hat{A}_j$ in Phase II of the minimum-cost procedure is

$$d_i^k = \tau_i - \alpha_i + \sum_{s=1}^{p} \hat{r}_{is}^k (\phi_s - \pi_s).$$

The possibility of employing alternative methods, i.e., alternative combinations of resources, affects this by allowing for a number of vectors $\hat{R}_i^k$. Thus to find the minimum $c_j - c_B B^{-1} \hat{A}_j$ one must find the minimum

$$\left[\sum_{A_i \in P_j} d_i^k \right],$$

over the permissible $\hat{R}_i^k$ as well as over all feasible combinations of arcs. The elements $\phi_s (s = 1, \ldots, p)$ are fixed for any problem, and the elements $\pi_s (s = 1, \ldots, p)$ are fixed for any iteration. One may, therefore, find the vector

$$\bar{R}_j^k = \hat{R}_i^k \epsilon t_{ik} \min \left[\sum_{s=1}^{p} \hat{r}_{is}^k (\phi_s - \pi_s) \right] \text{ for } i = 1, \ldots, m; k = 1, \ldots, q.$$

The kth matrix of these minima may then be defined as:

$$\bar{R}^k = [\bar{r}_{is}^k] (i = 1, \ldots, m; s = 1, \ldots, p).$$

Each column vector $\bar{R}_i^k = (\bar{r}_{i1}^k, \ldots, \bar{r}_{ip}^k)$ is the alternative combination of resources such that

$$\sum_{s=1}^{p} \bar{r}_{is}^k (\phi_s - \pi_s)$$

is minimized for arc i and commodity k. Now $\bar{R}^k$ may be substituted in the minimum-cost procedure previously discussed and the appropriate $\hat{A}_j$ selected for entry into the basis. Thus new $\bar{R}^k$ is constructed for each commodity, each iteration.

Summary of the Procedure

To summarize, the proposed procedure is:

1. Calculate $C_B B^{-1} = (\alpha_1, \ldots, \alpha_m, \pi_1, \ldots, \pi_p, \sigma_1, \ldots, \sigma_q)$.

2. For each arc-commodity pair, find the least-cost applicable resource vector, "cost" meaning cost in terms of the simplex multipliers and resources prices.

$$\bar{R}_i^k = \min_{\hat{R}_i^k \in t_{ik}} \left[\sum_{s=1}^{p} \hat{r}_{is}^k (\phi_s - \pi_s) \right].$$

3. For commodity $k = 1, \ldots, q$ calculate

$$d_i^k = \tau_i - \alpha + \sum_{s=1}^{p} \bar{r}_{is}^k (\phi_s - \pi_s) \text{ for } i = 1, \ldots, m,$$

and assign the d_i^k to the arc i as a pseudo cost.

4. Using the shortest-chain algorithm, find the chain with least

$$d_j^k = \sum_{i=1}^{m} a_{ij} \left[\tau_i - \alpha_i + \sum_{s=1}^{p} \bar{r}_{is}^k (\phi_s - \pi_s) \right] \text{ for } k = 1, \ldots, q.$$

5. Find

$$\min_k [d_j^k - \sigma_k].$$

6. If the minimum $[d_j^k - \sigma_k] < 0$, the vector $\hat{A}_j$ is entered in the basis. If $[d_j^k - \sigma_k] \geq 0$, there is no chain that may improve the value of the objective function, and the procedure is terminated.

Validity of the Procedure

Consider the linear programming formulation of the substitution problem. It is identical to the original cost-minimization problem except that every column vector in the original problem will be replaced by

$$\prod_{j \in p_j} N(M_i) \text{ [where } N(M_i) \text{ is the number of alternate resource vectors applying to arc } i\text{]},$$

alternate chains. The expanded substitution matrix will be many times larger than the original matrix, should either actually be enumerated.

It is claimed that the procedure outlined here will find the least-cost (in the sense previously described) vector to enter the basis. It should be noted that if the procedure does not find the least-cost vector, but some other vector, say the nth least-cost vector, the algorithm will progress toward an optimum solution in the early stages but will terminate early. That is, any vector that satisfies the simplex rule may be brought into the basis, but since the algorithm is terminated when the "shortest" chain does not satisfy the simplex rule, the validity of the procedure depends on the validity of the shortest-chain procedure.

Suppose that the chain produced as a candidate is not the shortest chain and there is some other candidate chain j^* for which

$$d^k_{j*} - \sigma_{k*} < d^k_j - \sigma_k.$$

Two possibilities for this other chain exist:

1. The shorter chain consists of the same arcs as our candidate chain, but has different (allowable) resource vectors associated with one or more of these arcs.

2. The shorter chain consists of different arcs altogether with some allowable set of resource vectors assigned to their respective arcs.

The first case is a chain that

$$d^k_{j*} < d^k_j$$

or that

$$\sum_{i=1}^{m} a_{ij}(\tau_i - \alpha_i) + \sum_{i=1}^{m} a_{ij} \sum_{s=1}^{p} r^{*k}_{is}(\phi_s - \pi_s) - \sigma_k$$

is less than

$$\sum_{i=1}^{m} a_{ij}(\tau_i - \alpha_i) + \sum_{i=1}^{m} a_{ij} \sum_{s=1}^{p} \bar{r}_{is}{}^k(\phi_s - \pi_s) - \sigma_k,$$

but since the first and last terms of each expression are identical, that is a claim that for at least one arc, common to both chains,

$$\sum_{s=1}^{p} r^{*k}_{is}(\phi_s - \pi_s) < \sum_{s=1}^{p} \bar{r}^k_{is}\ (\phi_s - \pi_s),$$

but since $\sum_{s=1}^{p} \bar{r}^k_{is}\ (\phi_s - \pi_s)$ $(i = 1, \ldots, m; k = 1, \ldots, q)$ is the minimum available (step 2), the claim that $d^k_{j*} - \sigma_{k*} < d^k_j - \sigma_k$ is inconsistent, and hence case 1 cannot occur. A true shortest chain must employ the least-cost allowable resources on each arc that is a member of the chain.

Case 2 resolves itself to a claim that there is some chain that uses the least-cost allowable resources on each of its member arcs and has a lower $(d^k_j{}_* - \sigma_{k*})$ than that of the candidate chain. Since the proposed procedure evaluates the pseudo cost of each arc incorporating the minimum resource costs $\left[\text{i.e.,} \sum_{s=1}^{p} \bar{r}^k_{is}(\phi_s - \pi_s)\right]$ and identical arc-use pseudo costs $\left[\text{i.e.,} \sum_{i=1}^{m} a_{ij}(\tau_i - \alpha_i)\right]$, a claim that case 2 exists is simply a claim that the shortest-chain algorithm does not find the shortest chain.

Usefulness of the Procedure

In order to be useful in application, the routes selected and resources assigned must be feasible in the object system. Routes through the network are composed of a series of arcs and the resources assigned to them. Again using a transportation network as an example, it is undesirable to have different vehicle types assigned to contiguous arcs of the same mode in a chain. That is, one wants the same vehicle to carry the commodity over all contiguous arcs of the same mode in a chain. Quarter-ton and 12-ton trucks may be feasible substitutes, but one does not wish to transfer from one to another at a node.

This is an important consideration if the results of the solution are to be used. It is simply not feasible in practice to use chains that employ different vehicles on various arcs of the same chain unless

the chain is multimode and transfer arcs are included. A procedure that is computationally simpler than the general method just described is available, and it guarantees that the same resource combinations will be used on all arcs of a chain that are of a particular mode.

A "master" resource vector representing the resources required to sustain a unit flow over a standard arc of unit length is provided for every commodity, mode, and method. Each arc then has a mode identifier, a condition factor, and a length factor assigned. The minimum-cost (in terms of the simplex multipliers) method is then selected for each iteration, and the resource vectors for each arc are generated using the condition and length scalars. Thus a single master vector, representing a particular method, is selected as the minimum-cost method for all arcs of that mode for each iteration. The solution may contain several chains from s_k to t_k each with arbitrarily different combinations of resources used, but each chain will be internally consistent with respect to resources used. Continuity of vehicle type is ensured for all chains in the solution.

COMPUTATIONAL EXPERIENCE

A computer program in FORTRAN IV for the Control Data 6400 has been developed for both maximum-flow and minimum-cost formulations incorporating the substitution feature. The program uses the product form of the inverse and will accommodate up to 150 commodities, 1,000 arcs, and 50 resources. Up to 20 modes are permitted and each mode may have up to three alternative resource-requirement vectors. Thus each arc may use any of three feasible combinations of resources to accomplish the move. A series of applications has been solved successfully and the results are encouraging with respect to accuracy and speed of solution. The use of the substitution feature does increase the time required for solution, but this increase has been small in the cases tested to date.

ACKNOWLEDGMENTS

The research leading to this note was done under contract with the Defense Communications Agency in support of the Special Assistant for Strategic Mobility, Joint Chiefs of Staff. We wish to acknowledge the encouragement and assistance given us by Mr. Donald Boyer, formerly at the George Washington University, Logistics Research Project, and Professor Mandell Bellmore of The Johns Hopkins University.

REFERENCES

[1] Allen, R. G. D., *Macro-Economic Theory* (St. Martin's Press, Inc., New York, 1968).

[2] Dantzig, G. B., and P. Wolfe, "The Decomposition Algorithm for Linear Programming," Econometrica, **29**, 767–78 (1961).

[3] Ford, L. R., Jr. and D. R. Fulkerson, "A Suggested Computation for Maximal Multi-Commodity Network Flows," Mgt. Sci. (Oct. 1958).

[4] Ford, L. R., Jr. and D. R. Fulkerson, *Flows in Networks* (Princeton University Press, Princeton, N. J., 1962).

[5] Hadley, G., *Linear Programming*, Addison-Wesley Publishing Co., Inc., Reading, Mass., 1962.

[6] Jewell, William S., "A Primal-Dual Multi-Commodity Flow Algorithm," ORC 66–24, Operations Research Center, University of California, Berkeley (Sept. 1966).

[7] Tomlin, J. A., "Minimum-Cost Multi-Commodity Network Flows," Operations Research, (Jan. 1966).

ADDITIONAL REFERENCES

Boyer, Donald D., "A Modified Simplex Algorithm for Solving the Multi-Commodity Maximum Flow Problem," TM-14930, The George Washington University Logistics Research Project, Washington, D.C. (Mar. 1968).

Busacker, R. G., et al., "Three General Network Flow Problems and Their Solutions," RAC-TP-183, Research Analysis Corporation (Nov. 1962).

Fitzpatrick, G. R., et al., "Programming the Procurement of Air Lift and Sealift Forces: A Linear Programming Model for Analysis of the Least Cost Mix of Strategic Deployment Systems," Nav. Res. Log. Quart. **14**, (1967).

Rao, M. R. and S. Zionts, "Allocation of Transportation Units to Alternative Trips—A Column Generation Scheme with Out-of-Kilter Subproblems," Operations Research (Jan.-Feb. 1968).

Sakarovitch, M., "The Multi-Commodity Maximum Flow Problem," Operations Research Center, University of California, Berkeley (Dec. 1966).

ON CONSTRAINT QUALIFICATIONS IN NONLINEAR PROGRAMMING

J. P. Evans
Graduate School of Business Administration
University of North Carolina

ABSTRACT

In this paper we examine the relationship between two constraint qualifications developed by Abadie and Arrow, Hurwicz, and Uzawa. A third constraint qualification is discussed and shown to be weaker than either of those mentioned above.

I. INTRODUCTION

In this paper we are concerned with constraint qualifications for the nonlinear programming problem

$$\begin{aligned} &(P)\\ &\min f(x)\\ &\text{s.t. } g_i(x) \leqq 0 \qquad i=1, \ldots, m, \end{aligned}$$

where f, g_i, $i=1, \ldots, m$, are real-valued functions defined on n-space. A constraint qualification, such as that of Kuhn and Tucker [6], places restrictions on the constraint functions of (P) such that if $x_0 \in E^n$ is an optimal solution for (P) and f and g_i, $i=1, \ldots, m$, are differentiable at x_0, then there exist scalars u_i, $i=1, \ldots, m$, satisfying the Kuhn-Tucker conditions*:

$$\nabla f(x_0) + \sum_{i=1}^{m} u_i \nabla g_i(x_0) = 0, \tag{1}$$

$$u_i g_i(x_0) = 0, \qquad i=1, \ldots, m, \tag{2}$$

$$u_i \geqq 0, \qquad i=1, \ldots, m. \tag{3}$$

Section II contains necessary background and notation. In Section II we also state the constraint qualifications of Arrow-Hurwicz-Uzawa [2] and the concept of sequential qualification due to Abadie [1]. Two examples then show that, although both of these qualifications are more general than that of Kuhn-Tucker [6], neither subsumes the other. In Section III we introduce the set of directions which are weakly tangent to a set and show that this concept leads to a weaker constraint qualification than either that of Arrow-Hurwicz-Uzawa or Abadie.

II. BACKGROUND AND NOTATION

For problem (P), let

$$S = \{x \mid g_i(x) \geqq 0, \qquad i=1, \ldots, m\};$$

*We denote the gradient of f evaluated at x_0 by $\nabla f(x_0)$; ∇f is considered to be a column vector ($n \times 1$).

and for $x_0 \in S$, let

$$I^0 = \{i | g_i(x_0) = 0\},$$

the set of effective constraints at x_0. Henceforth we will assume that f and g_i, $i \in I^0$, are differentiable at x_0. For completeness we now summarize relevant definitions from Abadie [1] and Arrow-Hurwicz-Uzawa [2].

DEFINITION 1: The linearizing cone at x_0 is the set of directions $X \in E^n$

$$C = \{X | X^T \nabla g_i(x_0) \leqq 0,\ i \in I^0\}.*$$

DEFINITION 2: A direction $X \in E^n$ is attainable at x_0 if there is an arc $x(\theta) \in E^n$, such that

(a) $x(0) = x_0$,

(b) $x(\theta) \in S$, $0 \leqq \theta \leqq 1$,

(c) $x'(0) = \lambda X$ for some scalar $\lambda > 0$.†

Now define

$$A = \{X | X \text{ is attainable at } x_0\}.$$

DEFINITION 3: A direction $X \in E^n$ is weakly attainable at x_0 if it is in the closure of the convex cone spanned by A.‡ Define

$$W = \{X | X \text{ is weakly attainable at } x_0\}.$$

DEFINITION 4: A direction $X \in E^n$ is tangent to S at x_0 if there exists a sequence $\{x^p\}$ in S such that $x^p \to x_0$ and a sequence $\{\lambda_p\}$ of nonnegative scalars such that

$$\lim_{p \to \infty} [\lambda_p (x^p - x_0)] = X.$$

Let

$$T = \{X | X \text{ is tangent to } S \text{ at } x_0\}.$$

Some properties of these sets are explored in [1] and [2]. Using these definitions we can summarize the constraint qualifications of interest.**

Kuhn-Tucker constraint qualification: $C \subseteq A$.††

(CQ) Arrow-Hurwicz-Uzawa constraint qualification: $C \subseteq W$.

(SQ) Abadie sequential qualification: $C \subseteq T$.

If any of the above conditions holds at the optimal point x_0, then conditions (1), (2), (3) have a solution (see [1], [2]).

By definition of the set W, it is clear that the Kuhn-Tucker constraint qualification implies (CQ). The following result establishes that condition (SQ) is implied by the Kuhn-Tucker qualification.

---

*This set is called the set of locally constrained direction by Arrow-Hurwicz-Uzawa [2]. The superscript T denotes transposition.

†$x'(0)$ denotes the derivative of the arc $x(\theta)$ at $\theta = 0$.

‡An example in [2] shows that A need not be closed.

**For convenience of reference we will denote the Arrow-Hurwicz-Uzawa qualification by (CQ) and that of Abadie by (SQ).

††The original statement of the Kuhn-Tucker constraint qualification involved the entire constraint set, S. In this note, as in [1] and [2], we are concerned with a local restriction which only need hold at the specific point x_0.

LEMMA 1: $A \subseteq T$.

PROOF: Suppose $X \epsilon A$; then there exists an arc $x(\theta)$ such that

$$\text{(a)} \quad x(0) = x_0,$$

$$\text{(b)} \quad x(\theta) \epsilon S, \quad 0 \leqq \theta \leqq 1,$$

$$\text{(c)} \quad x'(0) = \lambda X, \text{ some scalar } \lambda > 0.$$

Let $\{\theta_p\}$, $0 < \theta_p \leq 1$, be a sequence such that $\theta_p \to 0$. Define $\lambda_p = 1/\lambda\theta_p$, $p = 1, 2, \ldots$, and $x^p = x(\theta_p)$, $p = 1, 2, \ldots$. *Then* $x^p \to x_0$, and since $x(\theta)$ is differentiable at $\theta = 0$, we have

$$\lim_{p \to \infty} \lambda_p (x^p - x_0) = \lim_{p \to 0} (x^p - x_0)/\lambda\theta_p = X.$$

Thus $X \epsilon T$. Q.E.D.

The converse of Lemma 1 does not hold in general; see Example 2 below.

In the following examples we establish the lack of any ordering between (SQ) and (CQ).

EXAMPLE 1:

$$g_1(x) = \quad x_1 \quad x_2 \leq 0$$

$$g_2(x) = -x_1 \qquad \leq 0$$

$$g_3(x) = \qquad -x_2 \leq 0.$$

The constraint set, S, is the union of the nonnegative x_1- and x_2-axes. The following can be verified easily for $x_0 = \binom{0}{0}$:

$$C = \{X | X \geqq 0\}$$

$$A = S;$$

$$W = \{X | X \geqq 0\};$$

$$T = A.$$

Thus condition (CQ) holds, but (SQ) does not.

EXAMPLE 2: Define (following Abadie [1])

$$s(t) = \begin{cases} t^4 \sin 1/t & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}$$

$$c(t) = \begin{cases} t^4 \cos 1/t & \text{if } t \neq 0 \\ 0 & \text{if } t = 0. \end{cases}$$

As Abadie [1] observes, these functions are continuous with continuous first partial derivatives. The functions and the derivatives vanish at $t = 0$. Now consider

$$g_1(x) = x_2 - x_1^2 - s(x_1) \leqq 0$$

$$g_2(x) = -x_2 + x_1^2 + c(x_1) \leqq 0$$

$$g_3(x) = x_1^2 - 1 \leqq 0.$$

The set S is a collection of nonintersecting compact sets, one of which is the origin $\{\binom{0}{0}\}$. For $x_0=\binom{0}{0}$, we have

$$C=\{X|X_2=0\}=\text{the } x_1-\text{axis};$$
$$A=\{X|X=0\}=W;$$
$$T=C.$$

Thus condition (SQ) is satisfied, but (CQ) does not hold.

These two examples show that neither (SQ) nor (CQ) implies the other condition.

III. A NEW CONSTRAINT QUALIFICATION

The constraint qualification which we introduce in this section is a natural extension of the concept of tangents to a set used by Abadie [1].

DEFINITION 5: A direction $X\epsilon E^n$ is weakly tangent to S at x_0 if X can be written as a convex combination of tangents to S at x_0. Define

$$R=\{X|X \text{ is weakly tangent to } S \text{ at } x_0\}.*$$

In [1] (Lemma 3) it is shown that T is a closed nonempty cone; hence R is a closed convex cone. In the same paper it is shown (Lemma 4) that $T\subseteq C$. Since C is a closed convex cone and R is the convex cone generated by T, this establishes

LEMMA 2: $R\subseteq C$.

The constraint qualification of interest in this section can now be stated quite simply in terms of the sets C and R for the point x_0:†

(Q) $$C\subseteq R.$$

Since the set R is generated from the set T, it is clear that condition (SQ) implies (Q). In the remainder of this section we show that condition (CQ) implies (Q), and that if condition (Q) holds at x_0, and x_0 is optimal in problem (P), then the Kuhn-Tucker conditions hold at x_0.

LEMMA 3: Condition (CQ) implies condition (Q).

PROOF: Suppose X is a direction in C; since condition (CQ) holds, then $X\epsilon W$. W is the closure of the convex cone generated by A, and, by Lemma 1, $A\subseteq T$. Since T is closed, R, the convex cone generated by T, is also closed. Thus $W\subseteq R$, and condition (Q) holds. Q.E.D.

Lemma 3 together with the remarks preceding it establish that if either (CQ) or (SQ) holds at a point x_0, then (Q) holds there also.

THEOREM: Suppose f, g_i, $i=1, \ldots, m$ are differentiable at x_0, x_0 is optimal in problem (P), and $C\subseteq R$. Then there exist scalars u_i, $i=1, \ldots, m$, such that

(1) $$\nabla f(x_0)+\sum_{i=1}^{m} u_i \nabla g_i(x_0)=0$$

*Varaiya introduces the set R in [8] in a slightly different context.

†This qualification appeared in [4]; independently it appeared in a paper by Guignard [5], and subsequently in a footnote in Canon, Cullum, and Polak [3]. For completeness of the exposition we present a proof that this qualification is sufficient for the validity of the Kuhn-Tucker necessary conditions.

(2) $$u_i g_i(x_0) = 0, \quad i = 1, \ldots, m$$

(3) $$u_i \geqq 0 \qquad i = 1, \ldots, m.$$

The proof follows that of Abadie's Theorem 4 closely. We will employ the following version of Farkas' lemma.* Of the two linear systems

(I)	(II)†
$Au = b$	$x^T A \leqq 0$
$u \geqq 0$	$x^T b > 0,$

one and only one has a solution.**

PROOF: Now suppose (1), (2), and (3) have no solution. Then the system

$$u_i \geqq 0, \ i \in I^0$$

$$\nabla f(x_0) + \sum_{i \in I^0} u_i \nabla g_i(x_0) = 0$$

has no solution. But then by Farkas' lemma (identifying ∇f with $-b$ and ∇g_i, $i \in I^0$, with A), there is a direction $X \in E^n$, such that

(4) $$X^T \nabla f(x_0) < 0$$

$$X^T \nabla g_i(x_0) \leqq 0, \qquad i \in I^0.$$

Hence $X \in C$, the linearizing cone at x_0. Since $C \subseteq R$ and T is closed, X can be written as a convex combination of elements of T. That is for some collection $\{X^1, \ldots, X^k\} \subset T$, we have

(5) $$\begin{cases} \sum_{j=1}^{k} \eta_j X^j = X \\ \sum_{j=1}^{k} \eta_j = 1 \\ \eta_j \geqq 0 \qquad j = 1, \ldots, k. \end{cases}$$

Since $X^j \in T$, $j = 1, \ldots, k$, there exist sequences $\{x^{j,p}\} \subset S$ such that

$$\lim_{p \to \infty} x^{j,p} = x_0, \ j = 1, \ldots, k$$

and sequences $\{\lambda_{j,p}\}$ of nonnegative scalars such that

$$\lim_{p \to \infty} [\lambda_{j,p}(x^{j,p} - x_0)] = X^j, \ j = 1, \ldots, k.$$

Now for each $j = 1, \ldots, k$, by the differentiability of f at x_0, we have

$$f(x^{j,p}) = f(x_0) + (x^{j,p} - x_0))^T \nabla f(x_0) + \|x^{j,p} - x_0\| \epsilon_j,$$

where ϵ_j is a scalar which depends on p and j, and $\epsilon_j \to 0$ as $p \to \infty$ for each j. Thus for $j = 1, \ldots, k$,

*See Mangasarian [7].

†$x \geqq 0$ means $x_j \geqq 0$, $j = 1, \ldots, n$; $x \geqq 0$ means $x_j \geqq 0, j = 1, \ldots, n$ and $x_j > 0$ for at least one j.

**This is called the Second Transposition Theorem in Abadie [1].

$$(6) \qquad (f(x^{j,p}) - f(x_0))\lambda_{j,p} = \lambda_{j,p}(x^{j,p} - x_0)^T \nabla f(x_0) + \|\lambda_{j,p}(x^{j,p} - x_0)\|\epsilon_j.$$

In (6) multiply the jth equation by η_j and sum over $j = 1, \ldots, k$. Then

$$(7) \qquad \sum_{j=1}^{k} \eta_j (f(x^{j,p}) - f(x_0))\lambda_{j,p} = \left(\sum_{j=1}^{k} \eta_j \lambda_{j,p}(x^{j,p} - x_0)^T\right)\nabla f(x_0) + \sum_{j=1}^{k} \eta_j \|\lambda_{j,p}(x^{j,p} - x_0)\|\epsilon_j.$$

Now since X^j, $j = 1, \ldots, k$, is tangent to S at x_0, for sufficiently large p the right-hand side of (7) has the sign of $X^T \nabla f(x_0)$ which by (4) is negative. Thus for large enough p

$$\sum_{j=1}^{k} \eta_j \lambda_{j,p}(f(x^{j,p}) - f(x_0)) < 0.$$

But $\eta_j \geqq 0$, $j = 1, \ldots, k$, and $\lambda_{j,p} \geqq 0$ for each p and j. Thus for some j and p

$$f(x^{j,p}) < f(x_0).$$

Recalling that if X^j is tangent to S at x_0, then $x^{j,p} \in S$ for each $p = 1, 2, \ldots$, yields a contradiction of the optimality of x_0. Thus the Kuhn-Tucker conditions ((1), (2), (3)) have a solution at x_0. Q.E.D.

By an appropriate combination of the features of Examples 1 and 2 a case can be constructed for which condition (Q) holds, but neither of the qualifications (CQ) or (SQ) hold.

ACKNOWLEDGEMENTS

The author wishes to express his appreciation to David Rubin and F. J. Gould for helpful comments on this paper.

REFERENCES

[1] Abadie, J., "On the Kuhn-Tucker Theorem," *Nonlinear Programming*, edited by J. Abadie (John Wiley and Sons, Inc., New York, 1967).

[2] Arrow, K. J., L. Hurwicz, and H. Uzawa, "Constraint Qualifications in Maximization Problems," Nav. Res. Log. Quart. **8**, 175–191 (1961).

[3] Canon, M. D., C. D. Cullum, and E. Polak, *Theory of Optimal Control and Mathematical Programming* (McGraw-Hill Book Co., Inc., New York, 1970).

[4] Evans, J. P., "A Note on Constraint Qualifications in Nonlinear Programming," "*Center for Mathematical Studies in Business and Economics*," University of Chicago, Report 6917 (May 1969).

[5] Guignard, M., "Generalized Kuhn-Tucker Conditions for Mathematical Programming Problems in a Banach Space," *SIAM J. Control* **7**, 1969.

[6] Kuhn, H. W. and A. W. Tucker, "Nonlinear Programming," *Proceedings Second Berkeley Symposium* (University of California Press, Berkeley, 1951).

[7] Mangasarian, O., *Nonlinear Programming* (McGraw-Hill Book Co., Inc., New York, 1969).

[8] Varaiya, P. P., "Nonlinear Programming in Banach Space," *SIAM J. Appl. Math.* **15**, 284–293 (1967).

INVENTORY SYSTEMS WITH IMPERFECT DEMAND INFORMATION*

Richard C. Morey

Decision Studies Group

ABSTRACT

An inventory system is described in which demand information may be incorrectly transmitted from the field to the stocking point. The stocking point employs a forwarding policy which attempts to send out to the field a quantity which, in general, is some function of the observed demand. The optimal ordering rules for the general n-period problem and the steady state case are derived. In addition orderings of the actual reorder points as functions of the errors are presented, as well as some useful economic interpretations and numerical illustrations.

1. INTRODUCTION AND SUMMARY

Standard inventory models assume stochastic demands governed by known distribution functions. Superimposed on this inventory process is a known cost structure relative to which an optimal ordering policy is sought. Implicit in these models is the assumption that the demands are always accurately transmitted to the inventory stocking point. In practice, however, this assumption is frequently violated due to a variety of reasons which include improper preparation of requisitions, errors in keypunching and errors in transmission of data. The main effect of these errors is that the supply point may process a demand for an item which differs considerably from the true demand. This will, of course, increase the cost of an n-period model, say, and will lead to a different ordering policy. A study of the increased costs as a function of the variability of these errors would permit a rational evaluation of the effect of these errors.

Little research has been carried out on problems involving errors in inventory systems. Levy [4], [5] and Gluss [1] have published papers dealing with the general problem area, but from the standpoint of inexact estimates of the discount rate, penalty cost, and other constant parameters. Karlin [3] has studied inventory models in which the distribution of the demands may change from one period to another and obtains qualitative results describing the variation of critical numbers over time. Iglehart and Morey [2] have studied multiechelon systems in which optimal stocking policies are derived for the situation in which demand forecasts are used. In contrast, the problem suggested here deals with errors in the flow of real-time information from the demand point to the stocking point.

Our model will consider a single commodity. A sequence of ordering decisions is to be made periodically, for example, at the beginning of each quarter. These decisions may result in a replenishment of the inventory of the commodity. Consumption during the intervals between ordering decisions may cause a depletion of the inventory. The true demand in the field in each period is assumed to be a

*This research was supported by the Office of Naval Research, Contract Nonr-4457(00) (NR 347-001).

random variable, ξ, with a known distribution function. In addition, the transmitted demand back at the stocking point in each period is a different random variable, say η. Any differences between these two random variables arise due to human and mechanical shortcomings in the transmission of the demand. Finally, the stocking point employs a forwarding policy which attempts to send out to the field a quantity, say $g(\eta)$, which is, in general, a function of the observed demand, η. Since the stocking point is further constrained by the amount it has on hand, y, it forwards the smaller of y and $g(\eta)$. Figure 1 illustrates the flow of information and the flow of the stock. The dashed lines denote flow of information, and solid lines the flow of stock.

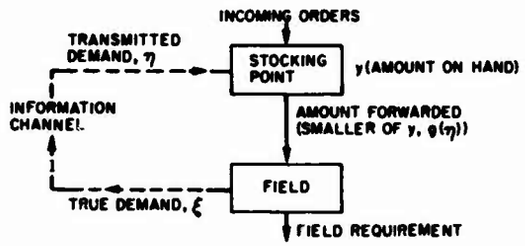


FIGURE 1. Information and Stock Flow.

The following costs are incurred during each period: a purchase or ordering cost $c(z)$, where z is the amount purchased; a holding cost $h(x)$, associated with the cumulative excess of supply over transmitted demand, which is charged at the end of the period; a shortage or penalty cost $p(x)$, associated with the excess of the true demand over the amount actually forwarded, which is also charged at the end of the period; and finally a salvage cost $r(x)$, which is associated with the excess of the forwarded amount over the true demand and can be interpreted as a credit or a revenue factor. Hence, the cost structure differs from the classical model in that the stocking point may forward *more* than what is actually desired in the field. In addition, although the penalties are still based on the difference between the amounts desired and the amounts forwarded, the amount forwarded is no longer limited solely by the amount on hand, but rather also by the amount which is thought to be desired. Throughout this paper we shall also assume that it is less costly to make purchases than to incur any shortages. Trivially, the optimal policy in the other case would be to never make any purchases.

The paper is organized as follows: We state and prove in Section 2 some general theorems which will subsequently be applied in Section 3 to various loss functions. These results permit qualitative orderings of the critical reorder points as functions of the particular demands and losses involved.

Section 3 is concerned with a more detailed discussion of the particular loss function arising naturally from an explicit consideration of the errors in the transmission of the demands. In this section, the general results of the previous sections are applied and various useful economic relationships and interpretations obtained.

Finally, Section 4 calculates numerically for some special cases the actual impact on the inventory system costs of various demand errors. Several strategies and their resulting costs are compared as a function of the standard deviation of the transmitted demand and as a function of the correlation between the true and transmitted demand.

2. CRITICAL NUMBERS FOR ORDERED LOSS FUNCTIONS

In this section we prove several theorems which relate critical reordering numbers to both the ordering of the loss functions and to the demands. These results will be applied in Section 3 to the

particular loss functions arising from a consideration of the errors in the transmission of the demands.

We first consider the one period case in which the ordering cost includes a set-up cost $K > 0$. We take

$$c(z) = \begin{cases} K + c \cdot z, & z > 0 \\ 0, & z = 0. \end{cases}$$

For the standard one period model with a convex L function, the optimal policy is of (s_1, S_1) type, where S_1 is the root of $c + L'(y) = 0$, and $s_1 < S_1$ satisfies

$$c \cdot s_1 + L(s_1) = K + c \cdot S_1 + L(S_1).$$

In our situation we wish to consider two one-period expected holding and shortage costs, L_1 and L_2. Let the corresponding optimal policies be $[s_1(i), S_1(i)]$, $i = 1, 2$. Then we easily obtain Theorem 1.

THEOREM 1: If $L_1'(y) \geq L_2'(y)$ for all real y, then $(i) S_1(1) \leq S_1(2)$ and $(ii) s_1(1) \leq s_1(2)$.

PROOF: Let $G_1(y; i) = c \cdot y + L_i(y)$, $i = 1, 2$. Then by our hypothesis, $G_1'(y; 1) \geq G_1'(y; 2)$ which implies (i). Define $s_1'(2) \leq S_1(1)$ as the solution to

$$G_1[s_1'(2); 2] = G_1[S_1(1); 2] + K.$$

Then by definition of $s_1(1)$, we have the equation

$$\int_{s_1'(2)}^{S_1(1)} G_1'(y; 2)\, dy = \int_{s_1(1)}^{S_1(1)} G_1'(y; 1)\, dy.$$

But since $G_1'(y; 2) \leq G_1'(y; 1) \leq 0$ for $y \leq S_1(1)$, we know that $s_1'(2) \geq s_1(1)$. But since $S_1(1) \leq S_1(2)$, we have $s_1'(2) \leq s_1(2)$ which yields $s_1(1) \leq s_1(2)$.

Assume now, that we have two inventory systems. The one-period expected costs (exclusive of ordering cost) are L_1 and L_2. The amounts of stock demanded from the inventory each period for the two systems are ξ_1 and ξ_2; with density functions ϕ_1 and ϕ_2, respectively. Making the usual assumption that L_1 and L_2 are convex, that the ordering cost for both systems is linear with unit cost $c > 0$, and that excess demand is completely backlogged, the optimal ordering policy for an n period is to order $[\bar{x}_n(i) - x]^+$ $(i = 1, 2)$. If we let $C_n(x; i)$ be the optimal expected cost for an n period model starting with initial inventory x; then $\bar{x}_n(i)$ is the smallest root of the equation $G_n'(x; i) = 0$, where

$$G_n(x; i) = c \cdot x + L_i(x) + \alpha \int_0^\infty C_{n-1}(x - \xi; i) \phi_i(\xi)\, d\xi.$$

Our next result requires a stochastic ordering of the demands ξ_1 and ξ_2. A random variable ξ_1 is stochastically less than ξ_2, written $\xi_1 < \xi_2$, if $\Phi_1(y) \geq \Phi_2(y)$ for all y, where Φ_i is the distribution function of ξ_i. The proof of Theorems 2 and 3 are direct extensions of Karlin [3] in that they permit an ordering of the critical reorder points as a function both of the ordering of the demands, and of the loss functions. The details of the proofs are therefore omitted.

THEOREM 2: If $L_1'(y) \geq L_2'(y)$ for all real y and $\xi_1 < \xi_2$, then

(*i*) $C_n'(x;\,1) \geq C_n'(x;\,2)$

and

(*ii*) $\bar{x}_n(1) \leq \bar{x}_n(2)$ for all $n \geq 1$.

The above result assumes complete backlogging. The next result generalizes Theorem 2 to include the case of lost sales.

THEOREM 3: If $L_1'(y) - c\phi_1(y) \geq L_2'(y) - c\phi_2(y)$, $\xi_1 < \xi_2$ and there is no backlogging of excess demands, then

(*i*) $C_n'(x;\,1) \geq C_n'(x;\,2)$ for $x \geq 0$

and

(*ii*) $\bar{x}_n(1) \leq \bar{x}_n(2)$ for all $n \geq 1$,

where

$\bar{x}_n(i)$ is the smallest root of the equation

$$c + L_i'(y) + \alpha \int_0^y C_n'(y-\xi;\,i)\phi_i(\xi)\,d\xi = 0.$$

Consider now a fixed time lag of λ periods ($\lambda \geq 1$) for delivery of ordered items and complete backlogging. Define

$$L_i^{(0)}(y) = L_i(y)$$

and

$$L_i^{(j)}(y) = \alpha \int_0^\infty L_i^{(j-1)}(y-\xi)\phi_i(\xi)\,d\xi,\ j \geq 1.$$

It is well known in inventory theory that the optimal ordering policy in the case of time lags is governed by a functional equation of the same type applicable in the case $\lambda = 0$, except that $L_i^{(\lambda)}(y)$ replaces $L_i(y)$. So, to obtain a result like Theorem 2 when $\lambda \geq 1$, we need only demonstrate the following result:

LEMMA 1: If $L_1'(y) \geq L_2'(y)$ for all real y, and $\xi_1 < \xi_2$, then

$$\frac{dL_1^{(j)}(y)}{dy} \geq \frac{dL_2^{(j)}(y)}{dy}$$

for all real y, and $j = 0, 1, 2 \ldots$.

PROOF: The result is true for $j = 0$ by hypothesis. Assume that it is true for $j-1$. Then

$$\frac{dL_1^{(j)}(y)}{dy} \geq \alpha \int_0^x \frac{dL_2^{(j-1)}}{dy}(y-\xi)\phi_1(\xi)d\xi = \frac{\alpha dL_2^{(j-1)}}{dy}(y-\xi)\Phi_1(\xi)\Big|_0^x + \alpha \int_0^x \frac{d^2L_2^{(j-1)}}{dy^2}(y-\xi)\Phi_1(\xi)d\xi,$$

but since

$$\Phi i(0)=0, \quad \Phi i(\infty)=1, \quad \frac{d^2L_2^{(j-1)}}{dy^2}(x) \geq 0 \text{ and}$$

$\Phi_1(\xi) \geq \Phi_2(\xi)$ for all ξ, it follows that

$$\frac{dL_1^{(j)}(y)}{dy} \geq \frac{dL_2^{(j)}(y)}{dy} \text{ for all } y.$$

3. A LOSS FUNCTION ARISING FROM CONSIDERATION OF ERRORED DEMANDS

In this section, we shall examine in detail a particular loss function arising from discrepancies in the true and transmitted demand.

The overall objective of this section is to develop qualitative results describing the variation of the critical numbers over time as a function of the forwarding poligy $g(\eta)$. Therefore, we will be primarily concerned with investigating functional relationships between the case of no errors and various treatments of the errored case. We shall also assume in what follows that the holding, shortage, and salvage cost are all linear. This assumption, while preserving the basic structure of the model, greatly facilitates the proofs and economic interpretations.

With this simplification, we find the loss function arising from employing a forwarding policy which attempts to send out to the field an amoung $g(\eta)$, whenever it observes at the stocking point a demand of η, is given by

$$Lg(y) = \{Eh\cdot[y-g(\eta)]^+\} + E\{p\cdot[\xi - y\wedge g(\eta)]^+\} - E\{r\cdot[y\wedge g(\eta)-\xi]\cdot\}.$$

Here, x^+ is x if $x \geq 0$, and 0 if x is less than 0, $a \wedge b$ denotes the smaller of a and b, and y denotes the inventory on hand at the stocking point at the beginning of a period after an order is received.

It should he noted in the case in which the true and transmitted demands are identical, and $g(\eta)=\eta$, that $L_g(y)$ reduces properly to the classical no error loss function.

To facilitate the investigation of $L_g(y)$, it will be convenient to define

$$1-Q(x) = Pr(\xi \geq x \text{ and } g(\eta) \geq x)$$

and

$$1-G(x) = Pr(g(\eta) \geq x),$$

where it will also be assumed that $G(0)=Q(0)=0$. Then it can be shown that

(1) $$L_g(y) = pE(\xi) + h\cdot y - (h+r)\int_0^y [1-G(u)]du - (p-r)\cdot\int_0^y [1-Q(u)]du,$$

(2) $$L_g'(y) = h-(h+r)[1-G(y)]-(p-r)[1-Q(y)],$$

and

(3) $$L_g''(y) = (h+r)G'(y) + (p-r)Q'(y).$$

Observe from expression 3 that a sufficient condition for $L_g(y)$ to be convex is that $p \geq r$: this will generally be the case since typically $p \geq c \geq r$.

Recalling also that the optimal steady state reorder level, call it $\tilde{x}(g)$, is the solution of $L_g'(y)=0$, it is clear that $\tilde{x}(g)$ is positive. This follows since $L_g'(0)=-p$, and $L_g'(\infty)=h$. Also it is interesting to observe from (2) that the optimal steady state reorder point increases as r, the salvage credit, increases. This result agrees with our intuition since we feel more disposed to keeping larger amounts of stock on hand (and hence, being in a position to send out larger quantities), if we can recover more of the amount by which we exceed the desired request.

It will be convenient to rewrite (2) as follows:

$$L_g'(y)=h[A_g(y)+D_g(y)]-pC_g(y)-rB_g(y)$$

where

$$B_g(y)=Pr[\xi \leq y \leq g(\eta)] \quad A_g(y)=Pr[g(\eta) \leq y \leq \xi]$$

and

$$C_g(y)=Pr[g(\eta) \geq y;\ \xi \geq y], \text{ and } D_g(y)=Pr[g(\eta) \leq y;\ \xi \leq y].$$

Now, define $L(y;\ \nu)$ to be the classical single period loss function if the demand is represented by the random variable ν. Then, substituting $\xi=\nu=g(\nu)$, in expression (2), we obtain

$$L(y;\nu)=E(h\cdot[y-\nu]^+)+E(p\cdot(\nu-y)^+),$$

and

$$L'(y;\nu)=h-(p+h)\cdot Pr(\nu \geq y).$$

Note that the oversupply credit factor r is not needed since in the classical formulation the stocking point never forwards to the field more than that which is actually desired.

Then the following qualitative relationships are available which will provide comparisons of the critical reordering levels for the perfect information case, and for various treatments of the situation in which transmission errors are present.

THEOREM 4:

(a) If $g(\eta)<\eta$, then $L_g'(y) \geq L'(y;\ \eta)$ for all y.

(b) If $g(\eta)<\xi$, then $L_g'(y) \geq L'(y;\ \xi)$ for all y.

PROOF: Since $p \geq r$,

$$L_g'(y)=h\cdot[A_g(y)+D_g(y)]-pC_g(y)-rB_g \geq h\cdot[A_g(y)+D_g(y)]-p\cdot[C_g(y)+B_g(y)],$$

but

$$A_g(y)+D_g(y)=Pr[\eta \leq g^{-1}(y)]=Pr[g(\eta) \leq y].$$

And since $g(\eta)<\eta$, we have $A_g(y)+D_g(y) \leq Pr(\eta \leq y)$ and $C_g(y)+B_g(y) \leq Pr(\eta \geq y)$.

Hence, $L_g'(y) \geq h\cdot Pr(\eta \leq y)-p\cdot Pr(\eta \geq y)=L'(y;\ \eta)$.

The proof of part (b) follows similarly.

Upon applying Theorems 1, 2, 3, and 4(a) we find that if the forwarding policy is to send out to the field an amount stochastically less than or equal to the transmitted demand, then the resulting critical numbers are always smaller than or equal to those obtained using the classical formula with a demand of η. This result is correct regardless of the distribution of the errors, regardless of the numbers of periods involved or delivery lag times, and finally, regardless of whether backordering or a lost sales philosophy is used. A similar interpretation can be given to Theorem 4(b). It is also noteworthy

to stress that the results do not depend upon having r greater than 0, and of course apply to the realistic forwarding policy of simply sending out, if possible, the transmitted demand η.

The next result provides a useful tool for determining, for any particular forwarding philosophy, how the steady-state critical numbers in the errored and perfect information cases compare.

LEMMA 2: Let $\tilde{x}$ denote the optimal steady-state reordering level with perfect information. Let $\tilde{x}(g)$ denote the optimal steady-state reordering level using a forwarding policy which sends out an amount $g(\eta)$ if the transmitted demand is η. Then, $\tilde{x}(g)$ is less than or greater than $\tilde{x}$ depending on whether

$$\frac{A_g(\tilde{x})}{B_g(\tilde{x})}$$

is larger than or smaller than the constant, $\dfrac{h+r}{h+p}$.

PROOF: It is easily shown that

$$L'(y;\xi)-L'_g(y)=(h+r)B_g(y)-(p+h)A_g(y).$$

Hence, since the steady-state solution satisfies $L'(y)=0$, the result follows directly.

Up to this point, we have assumed that knowledge of the transmitted demand was available before the decision had to be made as to the quantity to be sent out to the field. However, this is not always the case, especially in time of emergencies. The following result is useful in those important situations in which this transmitted demand information either is not available, or is of no value in forecasting the actual desired demand.

LEMMA 3: Assume the random variable ξ and η are independent. Then the optimal stationary forwarding policy is to send out to the field the constant amount

$$Q=F_{\xi}^{-1}\left(\frac{p-c}{p-r}\right)$$

and to reorder up to Q. In particular, the optimal fixed amount to be sent out in this situation is the βth quantile of F_{ξ} whenever $c=\beta r+(1-\beta)p$. The proof parallels directly the classical stationary single reorder level analysis and will be omitted.

4. NUMERICAL RESULTS

This section is concerned with attempting to isolate the cost or dollar consequences of errors in the demand for a particular case of practical interest. Two distinct types of analyses are presented. The first investigates how the inventory system costs vary as a function of the errors involved. Such knowledge is very useful in determining the amount of effort that should be spent to reduce the errors in the flow of demand information. Quite possibly in some situations the expense of eliminating the errors may be such that the savings resulting from having perfect information are not economically warranted.

Proceeding in a different spirit the second type of analysis is concerned with investigating the relative efficiencies of various stocking and forwarding strategies whose purpose it is to reduce the impact of the errors without requiring the costly elimination of the errors. Such strategies are definitely of interest due to the possibility that their use might enable a large portion of the costs currently associated with errors to be recouped with relativcly little additional effort.

Figure 2 depicts how the steady-state one period inventory system cost grows both as a function of p, the correlation coefficient between the true and transmitted demand, and as a function of σ_η, the standard deviation of the transmitted demand. The cost savings were computed assuming the joint true and transmitted demand are distributed according to a bivariate normal random variable (properly truncated to preserve the nonnegativity property) with equal means. This steady-state cost is computed using the loss function $L_\mu(\bar{x}_\eta)$ of expression (2), where $g(\eta)=\eta$, and $\bar{x}_\eta$ is the solution of $L'(y;\eta)=0$. Hence the difference between the dashed and solid line represents the actual incurred penalties resulting from naively using the classical reorder levels and are useful in determining to what degree it is economical to eliminate or reduce errors in the informational flow.

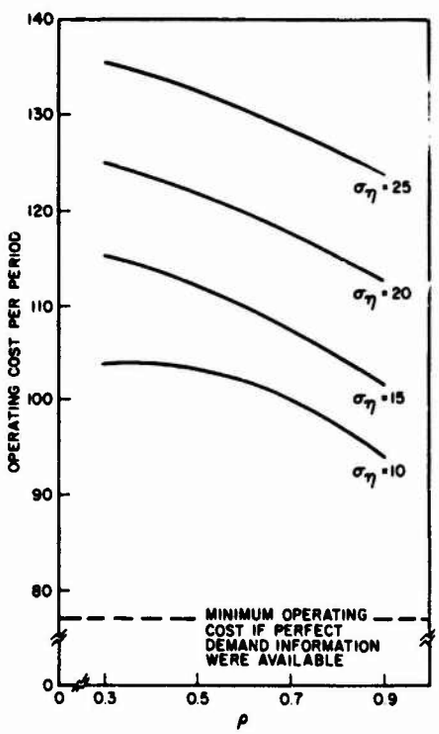


FIGURE 2. One period inventory systems cost as a function of the correlation coefficient with $c=1$, $h=0.25$, $r=0.30$, $p=5$, $E(\xi)=E(\eta)=75$, and $\sigma(\xi)=10$

The second type of analyses is concerned with the relative efficiencies of the following four forwarding strategies. In each case the optimal stocking policy for that particular forwarding policy was computed by solving $L_g'(y)=0$, and the corresponding cost calculated.

1. Send out the amount observed, i.e., $g(\eta)=\eta$.

2. Send out the optimal fixed amount, i.e., $g(\eta)=F_\xi^{-1}\left(\frac{p-c}{p-r}\right)$.

3. Send out the best estimate of the average demand, conditional upon the observation of the transmitted demand, i.e., $g(\eta)=E(\xi/\eta)$.

4. Send out a combination of strategy 2 and 3, i.e., $g(\eta)=F_{\xi/\eta}^{-1}\left(\frac{p-c}{p-r}\right)$.

In general, as might be expected, strategies 3 and 4 generally outperformed strategies 1 and 2. As proved earlier, the optimal stocking policy for strategy 1 resulted in lower reordering levels than those determined from solving $L'(y;\eta)=0$, and generally recovered about 10 percent of the cost due

to the errors. Strategy 2, while easy to implement, is obviously not efficient if there is a high correlation between ξ and η. Similarly, strategy 3, while having a certain heuristic appeal, does not perform as well as one might hope, mainly because it is independent of the various costs involved. On the other hand, the use of forwarding strategy 4, together with its appropriately derived ordering level, performed quite well and generally recouped from 50 to 58 percent of the costs incurred due to the errors in the demand. This is due clearly to strategy 4 being dependent both on the observed demand η as well as on the various inventory costs involved.

There is no doubt but that the implementation of some of these strategies would necessitate the use of extensive tables and probably would represent a realistic option only in case of a fully automated system. However it is felt that these and other strategies should continue to be investigated to the point where an economic balance can be achieved between the reduction of the errors on the one hand and the rational treatment of the remaining errors on the other.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Professor Donald L. Iglehart of Stanford University for helpful discussions on this model.

REFERENCES

[1] Gluss, Brian, "Cost of Incorrect Data in Optimal Inventory Computations," Management Science **6**, 491–495 (1960).

[2] Iglehart, D. and Morey, R., "Optimal Policies for a Multi-Echelon Inventory System with Demand Forecasts," forthcoming.

[3] Karlin, Samuel, "Dynamic Inventory Policy with Varying Stochastic Demands," Management Science **6**, 231–258 (1960).

[4] Levy, Joel, "Loss Resulting from the Use of Incorrect Data in Computing an Optimal Inventory Policy," Nav. Res. Log. Quart. **5**, 75–82 (1958).

[5] Levy, Joel, "Further Notes on the Loss Resulting from the Use of Incorrect Data in Computing an Optimal Inventory Policy," Nav. Res. Log. Quart. **6**, 25–32 (1959).

CONTRACT AWARD ANALYSIS BY MATHEMATICAL PROGRAMMING

Aharon Gavriel Beged-Dov

Weizmann Institute of Science, Rehovot, Israel and University of Toledo, Toledo, Ohio

ABSTRACT

A large manufacturer of telephone directories purchases about 100,000 tons of paper annually from several paper mills on the basis of competitive bids. The awards are subject to several constraints. The principal company constraint is that the paper must be purchased from at least three different suppliers. The principal external constraints are: 1) one large paper mill requires that if contracted to sell the company more than 50,000 tons of paper, it must be enabled to schedule production over the entire year; 2) the price of some bidders is based on the condition that their award must exceed a stipulated figure.

The paper shows that an optimal purchasing program corresponds to the solution of a model which, but for a few constraints, is a linear programming formulation with special structure. The complete model is solved by first transforming it into an almost transportation type problem and then applying several well-known L.P. techniques.

INTRODUCTION

This paper is based on a project directed by the writer on behalf of a large manufacturer of telephone directories. The company prints each year over 3,000 different directories in 20 printing plants across the nation. The individual directories, which vary in size from 50 to 2,000 pages, are printed in lots ranging from less than 1,000 up to 1,500,000 copies per year. For this purpose, the printers utilize paper in rolls of different widths. The roll widths, which range from 13 to 68 inches, depend upon the widths of the particular printing presses employed. The printers order the required paper from the Purchasing Organization of the company and Purchasing, in turn, distributes the orders among several paper mills, where the paper is manufactured in large reels ranging in width from 112 to 220 inches. Purchasing buys the paper on the basis of annual term contracts. The contracts are awarded in September, at which time both the requirements of the printers for the coming calendar year* and the terms of the paper manufacturers bidding for the business are known in detail.

The size of the individual awards depends on several factors. As a matter of policy, Purchasing strives to maintain multiple sources of supply. Specifically, at most, 40 percent of the total annual paper requirement of all the printers may be purchased from a single paper maker, regardless of how low his price may be.† Second, one major paper mill bids on the condition that if contracted to supply an amount of paper which exceeds half of his production capacity, he must be enabled to schedule production over the entire year, which, in slow periods, will tend to lower the awards to some of the

*In brief, the telephone directories are printed on presses which range from 11 to 68 inches in width and from 23 to 57 inches in circumference. Depending on the size of the press and the method of folding the printed sheets, a packet (known as a signature) which may contain from 24 up to 72 printed pages can be produced in a single revolution of the drum on which the text is mounted. Normally, the smaller the number of revolutions required to print a directory, the smaller the production cost. For example, a 720-page directory will be produced most economically on a 72-page signature press. Knowing, then, both the capacities of the presses he owns and the size of the different directories which he must print, each printer is able to determine how to schedule his presses most effectively, and, consequently, the amount of paper of a given width that he will require.

†See footnote on page 298.

other bidders. Third, the price of some bidders is based on the condition that they will be contracted to supply at least a given tonnage.

The typical contract obligates the company to purchase from a paper mill a specified amount of paper at a fixed unit cost, with provisions to compensate the supplier for excessive trim loss.* This means that the cost of ordering x tons of paper from supplier, s, for printer, p, cannot be less than $x(c_s + d_{sp})$ dollars. Here c_s the unit cost of the paper f.o.b. mill; d_{sp} is the unit transportation cost between the location of the seller and the location of the user. The cost may be higher, depending both on the actual trim loss incurred and the trim loss allowance stipulated in the contract. Suppose x^* is the trim loss incurred to fill the order, and α_s is the agreed allowance factor (usually 0.05). Then $c_{sp}(x)$, the total cost of the order, can be expressed as follows:

$$c_{sp}(x) = x(c_s + d_{sp}) + \lambda c_s(x^* - \alpha_s x),$$

where

$$\lambda = \begin{cases} 1 \text{ if } x^* > x\alpha_s \\ 0 \text{ otherwise.} \end{cases}$$

In principle, to minimize cost one only need to determine

$$c_{rp}(x) = \min_s c_{sp}(x),$$

and then order the paper from supplier r. However, x^* (and hence $c_{sp}(x)$) cannot be computed with sufficient accuracy at the time the allocation decision must be made since the manner in which the different suppliers will choose to trim the order from stock of reels of different widths they make is not known at this time. However, this difficulty can be resolved. Though it may be nearly impossible to forecast x^* accurately, the more general question of whether the stipulated trim allowance will, or will not, be exceeded can be answered. The reason is that the records of Purchasing show that throughout the years not a single request for additional payment has ever been submitted by a supplier. This means that the value of λ can be set to zero.

MATHEMATICAL FORMULATION

The fact that trim loss considerations may be ignored for the purpose of contract allocation makes it possible to formulate the problem of how to purchase the paper (required by the printers to print the directories assigned to them) economically as follows:

(A) $$\text{minimize: } Z = \sum_{s=1}^{S} \sum_{p=1}^{P} (c_s + d_{sp}) \sum_{k=1}^{K} x_{spk},$$

†The principal purpose of the policy is to increase availability. Clearly, should a strike or power breakdown, or other emergency occur in one place, at least part of the paper can still be obtained from the other. Should demand increase, it can be met with greater ease by calling upon the unused capacity of several, instead of only one or two, facilities. By the same token, the possibility of some paper mill becoming excessively dependent on the business, with the subtle responsibilities which such a position entails, is diminished. There are other advantages. A source of supply in close proximity to some printers may be secured, with a corresponding opportunity to save in transportation cost. Also, knowing that other companies are competing with him tends to keep each supplier alert to the needs of Purchasing. (Reference on page 297)

*The trim problem arises from the fact that the paper manufacturers must cut the large reels in which the paper is produced into smaller rolls of the widths ordered. Since the roll widths cut from a single reel rarely add perfectly to equal the reel width, a certain amount of paper at the edge of the reel is wasted. This waste, which is known as trim loss, is reflected in the cost of the paper.

suhject to

(A.1) $\sum_{p=1}^{P}\sum_{k=1}^{K} x_{spk} \leq M_s \qquad s = 1, 2, \ldots, S$

(A.2) $\sum_{p=1}^{P}\sum_{k=1}^{K} x_{spk} \leq Q_s \qquad s = 1, 2, \ldots, S,$

(A.3) $\sum_{s=1}^{S} x_{spk} = b_{pk} \qquad p = 1, 2, \ldots, P \quad k = 1, \ldots, K,$

(A.4) $\sum_{p=1}^{P}\sum_{k=1}^{K'} x_{spk} \leq \sum_{k=1}^{K'} r_{sk} \qquad s = 1, 2, \ldots, S \quad K' = 1, \ldots, K,$

(A.5) $\sum_{p=1}^{P}\sum_{k=1}^{K} x_{spk} \geq m_s \qquad \text{some } s,$

(A.6) $\sum_{p=1}^{P} x_{rpk} \leq \delta_r(A_r)\frac{t_k}{T} \qquad k = 1, 2, \ldots, K,$

(A.7) $\sum_{p=1}^{P} x_{rpk} \geq \delta'_r(A_r)\frac{t_k}{T} \qquad k = 1, 2, \ldots, K,$

(A.8) $x_{spk} \geq 0 \qquad \text{all } s, p, k.$

DEFINITION OF SYMBOLS

x_{spk} – the amount of paper shipped from supplier, s, to printer, p, for use in period, k.

b_{pk} – the amount of paper required by printer, p, in period, k.

M_s and m_s – the maximum and minimum awards which bidder s will accept.

Q_s – the maximum amount Purchasing will buy from him.

r_{sk} – the average production capacity of supplier, s, in period, k.

A_s – the amount of business actually awarded to supplier, s.

T and t_k – the annual and the periodic requirement for paper of all the printers.

r – the index of the supplier who insists on a continuous schedule.

δ_r and δ'_r – positive constants to be determined later. The formulation assumes S-suppliers, P-printers, and K-periods (of equal duration).

The objective function Z consists of SPK linearly additive terms. It represents the annual cost of supplying the printers with the paper required to print the telephone directories assigned to them. As shown, Z must be minimum, subject, of course, to the constraints stated. That a bidder cannot be contracted to supply an amount of paper which exceeds either the maximum quantity he is capable of making, or which Purchasing will buy from him is expressed in (A.1) and (A.2) respectively. That each printer must receive the paper he needs is stated in (A.3). Implicit here is the assumption that

$$\sum_{p=1}^{P} b_{pk} \leq \min\left(\sum_{s=1}^{S} r_{sk}, \sum_{s=1}^{S}\frac{Q_s}{K}, \sum_{s=1}^{S}\frac{M_s}{K}\right) k = 1, 2, \ldots, K.$$

Otherwise, of course, a feasible solution cannot exist. That the cumulative production capacity of a bidder must not be exceeded is stipulated in (A.4). That some bidders will not sell less than a specified minimum amount of paper is stated in (A.5). Finally, (A.6) and (A.7) provide, if there is a need, supplier, r, with a schedule which in any one period is proportional to the phased requirements of all the printers.

SOLVING THE MODEL

The above system of equations can be simplified considerably. To begin, Eqs. (A.1) and (A.2) can be readily combined into a single equation by letting

$$\sum_{p=1}^{P}\sum_{k=1}^{K} x_{spk} \leq a_s = \min\,(M_s, Q_s) \qquad s=1, 2, \ldots, S.$$

This, in turn, makes it possible to reduce Eqs. (A) to (A.3) into an ordinary transportation problem* with S origins and $KP+1$ destinations:

(B) $$\text{Minimize } Z=\sum_{i=1}^{S}\sum_{j=0}^{n} c_{ij}x_{ij},$$

subject to

(B.1) $$\sum_{j=0}^{n} x_{ij}=a_i \qquad i=1, 2, \ldots, S,$$

(B.2) $$\sum_{i=1}^{S} x_{ij}=b_j \qquad j=1, \ldots, n, \text{ and}$$

(B.3) $$x_{ij}\geq 0 \qquad \text{all } i, j,$$

where

$$b_0=\max\,(0, \sum_{i=1}^{S} a_i-\sum_{j=1}^{n} b_j),$$

$$c_{ij}=c_i+d_{ij} \qquad i=1, \ldots, S \qquad j=1, \ldots, n,$$

$$c_{i0}=0, \text{ and}$$

$$b_j=b_{pk} \qquad p=1, \ldots, P \qquad k=1, \ldots, K,$$

such that

$$p=j \text{ module } P \qquad \text{all } j$$

$$k=\begin{cases} j/P & \text{if } j/P \text{ is integer for all } j \\ \left[\dfrac{j}{P}\right]+1 & \text{otherwise, for all } j \end{cases}$$

$[N]$ meaning the largest integer contained in N.

By noting that a fictitious destination ($j=0$) has been added to drain the excess of supply over demand at zero unit cost, we can represent the tableau for this problem as in Table 1.

This transportation problem can readily and efficiently be solved with a special algorithm. In recognition of this fact the systems of Eqs. (B) to (B.3) will be referred to hereafter as the *favored problem.*

*The quantities M_s, Q_s and b_{pk} are assumed to be integers.

TABLE 1. *Transportation Tableau for Favored Problem*

Origin	b_0	b_1	b_2 . . . b_P	b_{P+1} . . . b_n	Availability
1	0	c_{11}	c_{12} . . . c_{1P}	c_{11} . . . c_{1P}	a_1
2	0	c_{21}	c_{22} . . . c_{2P}	c_{21} . . . c_{2P}	a_2
.	.	.	. . .	. . .	.
.	.	.	. . .	. . .	.
.	.	.	. . .	. . .	.
S	0	c_{S1}	c_{S2} . . . c_{SP}	c_{S1} . . . c_{SP}	a_S

In terms of the notation used to define the favored problem, the remaining constraints of the general award model are as follows:

$$\sum_{j=1}^{K'P} x_{ij} \leq \sum_{j=1}^{K'} r_{ij} \qquad i=1, \ldots, S \qquad K'=1, \ldots, K, \tag{B.4}$$

$$\sum_{j=1}^{n} x_{ij} \geq m_i \qquad \text{some } i, \tag{B.5}$$

$$x_{r,jP+1}+x_{r,jP+2}+ \ldots +x_{r,jP+P} \leq \delta_r \frac{t_{j+1}}{T} \; j=0, 1, \ldots, K-1, \tag{B.6}$$

$$x_{r,jP+1}+x_{r,jP+2}+ \ldots +x_{r,jP+P} \geq \delta'_r \frac{t_{j+1}}{T} \; j=0, 1, \ldots, K-1. \tag{B.7}$$

Now an approach suggested by the Method of Additional Restraints ([2], [5]) in which a smaller system (i.e., (B) to (B.3)) is solved first without regard to the other constraints (i.e., (B.4) to (B.7)) can be employed to solve the general award model. This approach offers an important computational advantage over procedures which work always with the complete system in a case where a priori considerations suggest that the solution of the smaller system, upon substitution, will satisfy the remaining constraints as well. Then, of course, the complete problem has been solved ([6], pp. 384–385). For example, it is easy to see that Eq. (B.4) is amenable to the method since M_i normally is nearly twice as large as a_i.

Constraint (B.5) can be handled by means of the method used to solve the (well-known) single price break problem of inventory theory ([4], pp. 238–241). To employ the method, let u be the index of a bidder who has placed a minimum award restriction, and assume that upon solving the favored problem, the result indicates that $A_u \geq m_u$. Then bidder u should be treated as if he had not placed the restriction in the first place. Suppose, however, that $A_u < m_u$. Let $Z_u(g)$ be the solution of a new favored problem such that the row corresponding with bidder u has been changed from

$$\sum_{j=1}^{n} x_{uj} \geq m_u \text{ to } \sum_{j=1}^{n} x_{uj} = g$$

and c_{u0} has been changed from zero to M, a very large positive number. Then, if $Z_u(m_u) < Z_u(0)$, the bidder should be awarded a contract for exactly m_u tons; otherwise, he should be awarded nothing. If b bidders are allocated originally less than their minimum, 2^b different combinations need to be examined in this manner.

Suppose then that this constraint has been taken care of and that the entire problem has been formulated in terms of the (remaining) active bidders only. Now, upon solving the resulting favored problem and examining the allocation awarded to supplier, r, it is possible to determine if he should be provided with a stable production schedule. There are two cases to consider. If A_r, the amount awarded to the supplier is less than half his production capacity, M_r, Eqs. (B.6) and (B.7) can be eliminated. Otherwise the values of δ_r and δ'_r must be stipulated to fix the bounds within which production will fluctuate throughout the contract year. From (B.6), it is easy to verify that δ_r cannot be smaller than A_r since

$$A_r = \sum_{j=1}^{n} x_{rj} \leq \delta_r \sum_{k=1}^{K} \frac{t_k}{T} = \delta_r.$$

Similarly, Eq. (B.7) indicates that δ'_r must not exceed A_r. Therefore, why not let

$$\delta_r = (1+\alpha)A_r, \qquad \delta'_r = (1-\alpha)A_r \qquad 0 \leq \alpha \leq 1.$$

As an example, let $t_1 = 100$, $t_2 = 95$, $t_3 = 110$, $A_r = 80$, and $\alpha = 0.1$. Then the shipments of mill r will be contained within 23 to 29 in period 1; 22 to 27 in period 2; and 26 to 32 in period 3.

At this stage of the analysis the favored system will contain m rows ($m \leq S$, depending on whether a supplier has been dropped or not) and $KP+1$ columns; Eq. (B.4) will contain mK rows and KP columns, and in the event that $A_r > 0.5M_r$, Eq. (B.6) and (B.7) will contain K rows and P columns.

Now the original award model is readily solved with an appropriate linear programming code. This is not recommended, however, since a more efficient and accurate solution method can be employed.

There are two cases to consider.

If $A_r < M_r/2$ the complete contract award problem can be expressed as follows:

(C) $$\text{Minimize } Z = cx + dy$$

subject to

(C.1) $$Fx = a$$

(C.2) $$Rx + Iy = r$$

$$x, y \geq 0,$$

where

$$c = (c_{10}, \ldots, c_{mn}), \; d = (0 \ldots, 0).$$

Here F is the matrix of coefficients of the favored problem, R is the matrix of coefficients of the system of equations dealing with the suppliers' capacity constraints, I is the unit matrix, and y is an $mKx1$ column vector. The elements of y are slack variables corresponding with the rows in R.

Suppose that upon substitution of x_0, a feasible optimal solution to the favored problem, inspection reveals that $Rx_0 + Iy = r$, $y \geq 0$. Then it is easy to show that x_0 is an optimal solution for the complete problem by noting that the optimality condition ([6], p. 244),

$$(w\ 0)\begin{bmatrix} F & 0 \\ R & I \end{bmatrix} \leqslant (c\ d),$$

where w is the vector of dual variables corresponding with the optimal solution, will be satisfied. On the other hand, if one or more of the elements of y are found to be negative, the Dual Simplex Algorithm can be initiated, but with some modification, as the algorithm requires "knowledge of an optimal, but not feasible solution to the primal, i.e., a solution to the dual constraints" ([7], p. 36). (An important advantage of the method is that the go-out vector is chosen before the come-in vector, so that if there is a choice, a go-out vector which does not alter the favored basis, say B, can be selected, with considerable reduction in computing effort.) The modification is necessary because B is not a square matrix, and consequently, the dual variables cannot be obtained in the usual manner. However, the current value of the dual variables can be determined readily as shown by Bakes [1].

On the other hand, if $A_r > M_r/2$ the larger problem:

(D) $$\text{Minimize } cx+dy+d_1y_1+d_2y_2+d_3y_3$$

subject to

$$\begin{bmatrix} F & 0 & 0 & 0 & 0 \\ R^+ & -1 & 0 & I & 0 \\ R^- & 0 & I & 0 & -I \end{bmatrix}\begin{bmatrix} X \\ y \\ y_1 \\ y_2 \\ y_3 \end{bmatrix}=\begin{bmatrix} a \\ r^+ \\ r^- \end{bmatrix}$$

$$x, y, y_1, y_2, y_3 \geqslant 0$$

need be considered. Here c and x are defined as before, and y, y_1, y_2, y_3, are defined as in Ref. [1], $d=d_1=(0, \ldots, 0)$, $d_2=d_3=(M, \ldots, M)$, R^+ is the matrix of coefficients of Eqs. (B.4) and (B.6), and R^- is the matrix of coefficients of Eq. (B.7). This formulation makes it always possible to construct, starting from B, a primal feasible basis for the complete problem, say B^*. That is, knowing B

$$B^*=\begin{bmatrix} B & 0 \\ R_B^* & 1^* \end{bmatrix}$$

can be determined by inspection. Here, analogous to the notation used in Eq. (C.2), R_B^* is made of the columns in R^+ and R^- which are continuations of the columns of B, and the elements of I^*, which can be determined by inspection, are either $+1$ or -1 in the main diagonal, and zero elsewhere. The basis will be optimal if

$$(ww^*)\begin{bmatrix} F & 0 & 0 \\ R^* & I & -I \end{bmatrix} \leqslant (cdd_1d_2d_3),$$

subject to

$$(ww^*)\begin{bmatrix} B & 0 \\ R_B^* & I^* \end{bmatrix} \leqslant (c_Bd^*),$$

where the elements of d^* are chosen from $(dd_1d_2d_3)$ to correspond with the columns of I^*. How to solve for (ww^*) and then, if necessary, proceed until the complete problem has been solved, is shown in detail in Ref. [1].

If desired, R^* can be further reduced in size. "If we feel that some secondary constraints which are not active for the optimal solution to the smaller problem will remain inactive, it is unnecessary to add them into the new basis" ([6], p. 400). Thus, so long as Q_i is much smaller than M_i, all i, the constraints on the cumulative production load of the suppliers can be omitted. Furthermore, Eq. (B.6) may also be deleted from R^* using a method of H. M. Wagner [9]. The method requires that for each deleted row, a new origin and a new destination are added to the favored system of equations in accordance with surprisingly simple rules. The advantage of the method resides in the fact that with available computer codes, very little additional time will be required to solve the larger favored problem. As an example, if both schemes are carried out, F will have $m + k$ rows and $m + k$ columns, but R^* will contain only K, instead of $(m + 2)K$ rows.

SOME RESULTS

The model described above was tested using actual data from a recent year. In that year, Purchasing bought 94,481 tons of paper at, as shown in Table 2, at a cost of $17,200,000. This cost was paid according to the (coded) price schedule shown in Table 3.

TABLE 2. *Monthly Paper Usage—Recent Year (Tons of Paper)*

Printer	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1	4,473	3,553	3,782	1,201	4,140	3,025	2,386	4,116	802	2,350	3,455	12	33,295
2	4,890	4,319	2,121	141	828	922	1,981	3,075	1,845	553			20,675
3	456	489	744	409	187	413	358	169	470	374	504	191	4,765
4	691	989	681	702	922	675	325	1,670	190	2,409	399	1,836	11,487
5	68	68	117	85	84	21	24	39	130	9	93	3	741
6	141	69	108	82	67	33	38	91	126	23	108	1	887
7	208	380	47	1,033	10	57	672	171	432	977	175	206	4,368
8	1,009	5	37	11	553	53	251	19	616	400	19	69	3,042
9	388	467	953	1,266		288	523	299	977	1,894	1,118		8,174
10	454	337	540	54	1,827	194	1,927	63	510	333	190	281	6,710
11	48	9	10	64	33	18	35	15	42	49	10	4	337
Total...	12,826	10,685	9,140	5,048	8,651	5,699	8,520	4,727	6,140	9,371	6,071	2,603	94,481

Upon solving the contract award formulation represented by the system of Eqs. (B) to (B.7), the allocation shown in Table 4 was obtained at a cost of $17,063,999. (To facilitate comparison, the figure also shows the actual awards assignment made by Purchasing for that year.) The values of the decision parameters employed in the solution are summarized in Table 5. It is important to note here that the very first solution of the favored problem yielded the optimal result, which testifies to the power of the Method of Additional Restraints.

The overall savings, about $142,000, clearly indicates the value of employing an assignment model in award analysis [8].

TABLE 3. *Price Schedule (per ton)*: $c_{ij} = c_i + d_{ij} - constant$

Printer / Mill	1	2	3	4	5	6	7	8	9	10	11
1	9.50	13.10	17.90	10.10	21.80	17.60	16.00	28.40	28.00	17.80	19.40
2	9.60	13.10	12.90	10.10	24.00	17.60	16.00	26.00	24.20	17.80	14.40
3	10.40	13.10	18.30	4.40	23.10	15.30	5.90	31.00	29.30	17.30	20.30
4	28.93	13.83	13.43	30.33	17.23	13.43	14.03	17.63	18.43	25.33	22.73
5	20.60	10.30	20.10	21.60	22.60	19.80	16.70	29.20	28.00	11.90	15.30

TABLE 4. *Values of Decision Parameters*

Mill / Parameter	1	2	3	4	5
m (tons)	30,000	–	10,000	–	–
M (tons)	100,000	20,000	40,000	35,000	43,000
Q (tons)	45,000	20,000	20,000	22,000	25,000
K (periods)	12	12	12	12	12

It is conceivable that efficient allocation could be obtained using some trial and error method of analysis. However, considering that the amount of directory paper purchased by the company is enormous and the cost of implementing the assignment model is practically nil (altogether, about 15 minutes of IBM 1620 Computer time, and about 2 hours of human time were expended to achieve the results reported in this section) there is little justification for not taking advantage of a tool which can yield maximum results at a minimum of cost and effort.

TABLE 5. *Model* vs *Actual Allocation* (*tons*)

Mill / Printer	1	2	3	4	5
1	33,295				
2					20,625
3		4,765			
4			11,487		
5				8,174	
6				6,710	
7				4,368	
8				887	
9				337	
10					3,042
11		741			
Model	33,295	5,506	11,487	20,476	23,717
Actual	40,901	19,521	16,687	9,191	8,181

REFERENCES

[1] Bakes, M. D., "Linear Programming Solution with Additional Constraints," Operations Research Quarterly, Vol. **XVII**, No. 4 (Dec. 1966), pp. 425–445.

[2] Beged-Dov, A. G., "Some Computational Aspects of the M Paper Mills and P Printers Paper Trim Problem," Journal of Business Administration, Vol. **11**, No. 2 (June 1970), pp. 1–21.

[3] Beged-Dov, A. G., "Optimal Assignment of R&D Projects In a Large Company Using an Integer Programming Model," IEEE Transactions on Engineering Management, Vol. EM-**12**, No. 4 (Dec. 1965), pp. 138–142.

[4] Churchman, C. W., R. L. Ackoff, and E. Arnoff, Introduction to Operations Research (John Wiley & Sons, New York, 1957).

[5] Dantzig, G. B., "Upper Bounds, Secondary Constraints and Block Triangularity," Econometrica, Vol. **XXIII** (1955), pp. 174–183.

[6] Hadley, G., Linear Programming (Addison Wesley Publishing Company, Inc., Reading, Mass., 1963).

[7] Lemke, C. E., Jr., "The Dual Method of Solving Linear Programming Problems," Nav. Res. Log. Quart., Vol. I, No. 1 (Mar. 1954), pp. 36–47.

[8] Stanley, E. D., D. P. Honig, and L. Gainen, "Linear Programming in Bid Evaluation," Nav. Res. Log. Quart., Vol. I, No. 1 (Mar. 1954), pp. 48–54.

[9] Wagner, H. M., "On a Class of Capacitated Transportation Problems," Management Science, Vol. V, No. 3, pp. 304–318.

A FINITENESS PROOF FOR MODIFIED DANTZIG CUTS IN INTEGER PROGRAMMING

V. J. Bowman, Jr.
Carnegie-Mellon University
and
G. L. Nemhauser
Cornell University

ABSTRACT

Let

$$x_i = y_{i0} - \sum_{j \in R} y_{ij} x_j, \; i = 0, \ldots, m$$

be a basic solution to the linear programming problem

$$\max x_0 = \Sigma_j c_j x_j$$

subject to: $\Sigma_j a_{ij} x_j = b_i, \; i = 1, \ldots, m,$

where R is the index set associated with the nonbasic variables. If all of the variables are constrained to be nonnegative integers and x_u is not an integer in the basic solution, the linear constraint

$$\sum_{j \in R_u^*} x_j \geq 1, \; R_u^* = \{j | j \in R \text{ and } y_{uj} \neq \text{integer}\}$$

is implied. We prove that including these "cuts" in a specified way yields a finite dual simplex algorithm for the pure integer programming problem. The relation of these modified Dantzig cuts to Gomory cuts is discussed.

Consider the pure integer programming problem

(1) $$\max x_0 = \sum_j c_j x_j$$

subject to: $$\sum_j a_{ij} x_j = b_i, \; i = 1, \ldots, m$$

$$x_j \geq 0 \text{ and integer}, \; j = 1, \ldots, n.$$

It is assumed that the c_j and a_{ij} are integers and that x_0 has both an upper and a lower bound.

Let $x_0, x_1, \ldots, x_m$ be basic (not necessarily feasible) variables and R be the index set associated with the nonbasic variables. Expressing the basic variables in terms of the nonbasic variables, we have

(2) $$x_i = y_{i0} - \sum_{j \in R} y_{ij} x_j, \; i = 0, \ldots, m.$$

Suppose at least one y_{i0} given by (2) is not an integer. Then the integer constraints imply the linear constraint (3), which is not satisfied by the current solution

$$\sum_{j\in R} x_j \geq 1. \tag{3}$$

Equation (3) is a Dantzig cut [2]; however, (1) implies tighter cuts of the kind in which all coefficients are $+1$. Specifically, suppose x_u is not an integer in the basic solution given by Eq. (1). Then, as noted by Charnes and Cooper [1], the requirement that x_u be an integer implies that

$$\sum_{j\in R'_u} x_j \geq 1, \tag{4}$$

where $R \supseteq R'_u = \{j | j\in R \text{ and } y_{uj} \neq 0\}$.
The cut of Eq. (4) can be sharpened still further by noting that

$$\sum_{j\in R^*_u} x_j \geq 1, \tag{5}$$

where $$R'_u \supseteq R^*_u = \{j | j\in R \text{ and } y_{uj} \neq \text{integer}\}$$

is also implied by the integer requirements.

Gomory and Hoffman [5] have proved that Dantzig cuts (Eq. (3)) are not sufficiently strong to guarantee convergence of a linear programming algorithm to an optimal integer solution. We will show that the tighter cuts, given by Eqs. (4) and (5), when included in a certain way, yield a finite dual simplex algorithm.

THE ALGORITHM

1. Using the objective function as the top row of the tableau, solve (1) ignoring the integer constraints. If the optimal solution obtained is all-integer, terminate; otherwise add the redundant inequality $\sum_{j\in R} x_j \leq M$ (M is positive and very large) as the second row of the tableau to insure that the columns are lexicographically positive. Then go to step 2.

2. Let row u be the topmost row in which the basic variable is noninteger. Adjoin the constraint (5) to the bottom of the tableau, i.e.

$$s - \sum_{j\in R^*_u} x_j = -1,$$

and execute one dual simplex iteration with the new row as the pivot row. The pivot column must be chosen to maintain lexicographically positive columns. If the solution is all-integer and primal feasible, terminate; otherwise go to step 3.

3. If the solution is primal feasible or if $y_{00}, \ldots, y_{u-1,0}$ are integers and y_{u0} has not decreased by at least its fractional part, go to step 2; otherwise go to step 4.

4. Execute dual simplex iterations in the usual manner (lexicographically positive columns must be maintained) until primal feasibility is attained. If the solution is all-integer, terminate; otherwise go to step 2.

The branching rule of step 3 can be modified in several ways without affecting convergence; however, we have not been able to prove that it is always possible to go to step 4 when there are primal infeasibilities.

FINITENESS PROOF

THEOREM: Application of the above algorithm to a pure integer programming problem as given by (1) yields an optimal solution after a finite number of dual simplex iterations.

PROOF:* Clearly, the number of pivots in step 1 is finite.

Let $y_{ij}(0)$ be the entries in the tableau associated with an optimal linear programming solution (the solution obtained from step 1) and $y_{ij}(t)$ the entries in the tableau after t dual simplex iterations beyond step 1. The $y_{00}(t)$ form a monotone nonincreasing sequence bounded from below. Let Δ be the greatest integer such that $y_{00}(t) \geq \Delta$ for all t. For some t suppose we have

$$y_{00}(t) = \Delta + f_{00}(t),\ 0 < f_{00}(t) < 1.$$

In step 2, we add the cut

$$s - \sum_{j \in R_0^*} x_j = -1.$$

In the transformed tableau, we have

$$y_{00}(t+1) = y_{00}(t) - y_{0k}(t),$$

where k is the pivot column and $y_{0k}(t) \leq f_{00}(t)$.

From the definition of R_0^* it follows that $y_{0k}(t) > 0$ and consequently $y_{00}(t+1) < y_{00}(t)$.

We now show that there exists a $T \geq t+1$ such that $y_{00}(t^*) = \Delta$ for all $t^* \geq T$. Let $f_{0j}(t)$ be the fractional part of $y_{0j}(t)$ and $f_{0j}(t) = e_{0j}(t)/D(t)$, where $D(t)$ is the absolute value of the product of all previous pivot elements. Note that, since the a_{ij} are integer, $D(t)$ and $e_{ij}(t)$ are integers. Since the pivot element is -1, $D(t+1) = D(t)$ and

$$\begin{aligned} y_{00}(t+1) &= \Delta + \frac{e_{00}(t) - e_{0k}(t)}{D(t)} \\ &\leq \Delta + \frac{e_{00}(t) - 1}{D(t)}. \end{aligned}$$

If $y_{00}(t+1) > \Delta$, we add another cut from the objective row and again reduce the value of the objective function by at least $1/D(t)$. Consequently, after at most $e_{00}(t)$ cuts have been added, the objective function reaches Δ. Since the columns are maintained lexicographically positive, a similar argument can be used to show that the remaining variables become integers in a finite number of iterations.

Discussion and Comparison with Gomory Cuts

The proof just given for cuts from Eq. (5) applies as well to the weaker cuts from Eq. (4), but not, of course, to the Dantzig cuts of Eq. (3). The cuts of Eqs. (4) and (5), when derived from the objective row, reduce $y_{00}(t)$ by at least $1/D(t)$. This reduction is crucial. The Dantzig cut, on the other hand, yields no reduction in $y_{00}(t)$ whenever there is dual degeneracy.

A Gomory [4] cut taken from Eq. (2), when x_u is not integer, is

$$\sum_{j \in R_u^*} f_{uj} x_j \geq f_{uo} \tag{6}$$

*We assume, for simplicity, that the constraint set of (1) contains at least one lattice point. An empty constraint set will be indicated by unboundedness in the dual problem.

where f_{ij} is the fractional part of y_{ij}. Thus, the cuts given by Eqs. (5) and (6) involve different inequalities on the same subset of nonbasic variables. The constraint (5) cuts equally deep into each axis of the variables associated with the index set R_u^*. The Gomory constraint (6) cuts a different amount into each axis, depending upon the fractional parts of the coefficients in Eq. (2).

One might argue that for a randomly selected row

$$Pr(f_{ij} \geq f_{i0}) = Pr(f_{ij} < f_{i0}) = 1/2,$$

so that on the average constraint (5) should do as well as constraint (6). However, a reason for believing that (6) is superior to (5) is that (6) can have f_{i0}/f_{ik} large (>1) for the pivot index k.

The finiteness proof for the cuts of Eqs. (4) and (5), when compared with the very similar proof for Gomory cuts, highlights this point. When a Gomory cut is taken from the objective row, the objective function decreases by at least its fractional part.

For the modified Dantzig cuts, only the much smaller decrease of $1/D(t)$ is assured. In fact, to prove finiteness, we had to add cuts in certain cases when there were primal infeasibilities (see step 3 of the algorithm) to prevent the product of the pivots from increasing. When using Gomory cuts, primal infeasibilities can always be removed (and therefore further reductions can be obtained in the objective) before additional cuts are made. Conceivably, the constraint of Eq. (4) may represent the weakest cut for which a finitely convergent linear programming process can be constructed.

There is a much closer relationship between cuts (5) and (6) than the mere fact they are linear inequalities on the same subset of nonbasic variables. Specifically, the cut of Eq. (5) is a linear combination of two Gomory cuts. Glover [3] has generalized the representation of Gomory cuts to yield cuts, from row u of Eq. (2), of the form*

$$\sum_{j \in R} (\langle h y_{uj} \rangle - \langle h \rangle y_{uj}) x_j \geq \langle h y_{u0} \rangle - \langle h \rangle y_{u0}, \tag{7}$$

where $\langle x \rangle$ denotes the least integer $\geq x$ and h must be chosen so that $\langle h y_{u0} \rangle - \langle h \rangle y_{u0} > 0$ (y_{u0} is not an integer).

Choosing the parameter h to be integer yields the finite abelian group of Gomory cuts of the method of integer forms [4]. In particular the cut of Eq. (6) is obtained with $h = -1$. Setting $h = +1$ yields

$$\sum_{j \in R_u^*} (1 - f_{uj}) x_j \geq 1 - f_{u0}. \tag{8}$$

*Glover actually uses the two parameter representation

$$(\langle h \rangle - p) x_u + \sum_{j \in R} (\langle h y_{uj} \rangle - p y_{uj}) x_j \geq \langle h y_{u0} \rangle - p y_{u0}.$$

If $\langle h \rangle - p$ is not zero, $x_u = y_0 - \sum_{j \in R} y_{uj} x_j$ must be substituted into the cut equation to obtain a basic solution. This substitution is equivalent to requiring $p = \langle h \rangle$. For practical purposes then, Glover's generalized cuts are one parameter. Many of Glover's arguments for deriving properties of these cuts can be simplified by setting $p = \langle h \rangle$. However, the use of p does emphasize that two different quantities $\langle h \rangle$ and h influence the nature of the cut.

Adding the two Gomory cuts of Eqs (6) and (8), we obtain

$$\sum_{j\in R_2^*}(1-f_{uj})x_j+\sum_{j\in R_1^*}f_{uj}\geq 1-f_{uo}+f_{uo}$$

which is precisely the cut of Eq. (5).

Finally, Glover has observed that the sum of two cuts taken from (7), one having $h=h_1$ and the other having $h=h_2$ with $\langle h_2\rangle=-\langle h_1\rangle$, yields a cut with integer coefficients. Such cuts can have all of the coefficients $=+1$ but with even fewer nonbasic variables appearing in the sum than in Eq. (5).

REFERENCES

[1] Charnes, A., and W. W. Cooper, "*Management Models and Industrial Applications of Linear Programming* (Wiley, New York, 1961), Vol. **II**, pp. 700–701.

[2] Dantzig, G. B., "Note on Solving Linear Programs in Integers," Nav. Res. Log. Quart. **6**, 75–76 (1959).

[3] Glover, F., "Generalized Cuts in Diophantine Programming," Man. Sci. **13**, 254–268 (1966).

[4] Gomory, R. E., "An Algorithm for Integer Solutions to Linear Programs," in *Recent Advances in Mathematical Programming*, edited by R. L. Graves and P. Wolfe (McGraw-Hill, New York, 1963), pp. 269–302.

[5] Gomory, R. E., and A. J. Hoffman, "On the Convergence of an Integer Programming Process," Nav. Res. Log. Quart. **10**, 121–124 (1963).

A SOLUTION FOR QUEUES WITH INSTANTANEOUS JOCKEYING AND OTHER CUSTOMER SELECTION RULES

R. L. Disney

Department of Industrial Engineering
The University of Michigan

and

W. E. Mitchell

Logistics Department
Standard Oil, New Jersey

ABSTRACT

This paper presents a general solution for the $M/M/r$ queue with instantaneous jockeying and $r > 1$ servers. The solution is obtained in matrices in closed form without recourse to the generating function arguments usually used. The solution requires the inversion of two $(2^r-1) \times (2^r-1)$ matrices.

The method proposed is extended to allow different queue selection preferences of arriving customers, balking of arrivals, jockeying preference rules, and queue dependent selection along with jockeying.

To illustrate the results, a problem previously published is studied to show how known results are obtained from the proposed general solution.

1.0 THE PROBLEM

1.1 Queue Selection Rules

We consider the following queueing situation. There are r servers. The probability distribution functions of service times are negative exponential distributions with parameters $\mu_1, \mu_2, \ldots, \mu_r$. Arrivals form a Poisson stream with parameter λ. Initially, we will assume that all customers that arrive will join a queue (see Section 5 for other queue behaviors). Each server is assumed to have his own waiting line. Arrivals will join a queue according to the following rules:

1.1(a) If all queues are empty, he will choose any of the open queues with equal probability.

1.1(b) If several, say c, but not all queues are empty, then an arrival will join any of the empty queues with probability $1/c$.

1.1(c) If all queues are occupied, then the customer will join the shortest queue.

1.1(d) If all queues are occupied, and if several queues, say $s \leq r$, have numbers in them equal to the number in the shortest queue, then the customer chooses any of these equally short queues with probability $1/s$.

1.2 Jockeying Rules

Once a customer has joined a queue, he will be allowed to change queue (jockey) in accordance with the following rules (see Section 5 for other jockey rules):

1.2(a) If, at any time, $n_i - n_j \geq 2$, then the customer in the ith queue will jockey to jth queue instantaneously.

1.2(b) If, under rule 1.2(a), it is possible for the customer in the ith queue to jockey to several queues, say s, then he will jockey to any of the eligible queues with probability $1/s$.

1.2(c) If for $r > 2$, $n_i - n_j \geq 2$ for fixed i and j and $n_h - n_j \geq 2$, then a jockey from i or h is equally likely.

This problem has been studied by Haight [2] and Koenigsberg [3] for $r = 2$. It is called a queueing system with instantaneous jockeying.

2.0 SYSTEMS EQUATIONS

2.1 The Transition Diagram

The simplest way to view this problem is to consider its transition diagram. Using the usual "steady state" arguments one can develop the "steady state" equation (if necessary) from this diagram. From the transition diagram it will be evident that the coefficient matrix for the "steady state" equations exhibit a considerable amount of regularity that can be exploited to solve the problem.

The random process of interest to us ("the number of customers at each server at time t") will have a state space consisting of r-tuples whose jth element gives the number of customers before server j. We define the vectors

$$\mathbf{n} = (n, n, \ldots n), n = 0, 1, 2, \ldots$$

$$\mathbf{n}_i = (n, n \ldots \underset{\substack{j\text{th} \\ \text{element}}}{n+1} \ldots n), n = 0, 1, 2, \ldots$$

$$\mathbf{n}_{ij\ldots h} = (n, n, \ldots \underset{\substack{i \\ \text{element}}}{n+1}, \ldots \underset{\substack{j \\ \text{element}}}{n+1} \ldots \underset{\substack{h \\ \text{element}}}{n+1} \ldots n), n = 0, 1, 2 \ldots .$$

$$i = 1, 2, \ldots r.$$

$$j > i$$

$$h > j$$

We define the following state probabilities for $r = 2$,

p_0 = probability that the system is in state $\mathbf{0}$

p_n = probability that the system is in state $\mathbf{n}$, $n = 1, 2, \ldots$

p_{n_1} = probability that the system is in state $\mathbf{n}_1$, $n = 0, 1, 2, \ldots$

p_{n_2} = probability that the system is in state $\mathbf{n}_2$, $n = 0, 1, 2, \ldots$

The transition diagram for $r = 2$ is given in Figure 1. We omit transitions from a state to itself. Such transitions do not contribute to our later work. Generalizations for $r > 2$ are obvious. The indicated transition rates follow simply from the queueing rules given in Sections 1.1 and 1.2. Thus, for example, if the state of the system is $\mathbf{1}_1$ a transition to $\mathbf{1}$ occurs either by server 1 completing service on one of the two customers in his queue (rate μ_1) or server 2 completing service on the only customer he has (rate μ_2). In the latter case a customer immediately jockeys from server 1 to put the system into state $\mathbf{1}$.

2.2 State Equations

Using the usual methods of equilibrium analysis one can write the steady state equation from Figure 1.

(a) For $\mathbf{0}$

$$-\lambda p_0 + \mu_1 p_{0_1} + \mu_2 p_{0_2} = 0$$

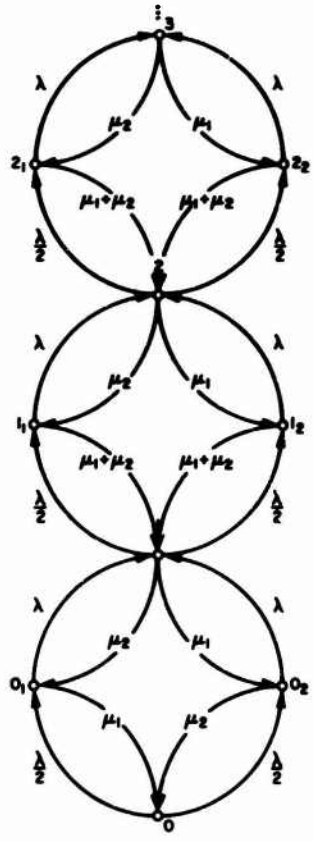


FIGURE 1. Transition diagram for $r = 2$

(b) For $\mathbf{0}_1$

$$\frac{\lambda}{2} p_{\mathbf{0}} - (\lambda + \mu_1)\, p_{\mathbf{0}_1} + \mu_2 p_1 = 0$$

(c) For $\mathbf{0}_2$

$$\frac{\lambda}{2} p_{\mathbf{0}} - (\lambda + \mu_2)\, p_{\mathbf{0}_2} + \mu_1 p_1 = 0.$$

For $\mathbf{n} > 0$ it is apparent that equations for $\mathbf{n}, \mathbf{n}_1, \mathbf{n}_2$ do not depend on the particular values of $\mathbf{n}$. Hence for each $\mathbf{n} > 0$ one has

(a) For $\mathbf{n+1}$,

$$\lambda(p_{\mathbf{n}_1} + p_{\mathbf{n}_2}) - (\lambda + \mu_1 + \mu_2) p_{\mathbf{n}+1} + (\mu_1 + \mu_2) p_{(\mathbf{n}+2)_1} + (\mu_1 + \mu_2) p_{(\mathbf{n}+2)_2} = 0$$

(b) For $(\mathbf{n+1})_1$

$$\frac{\lambda}{2} p_{\mathbf{n}+1} - (\lambda + \mu_1 + \mu_2) p_{(\mathbf{n}+1)_1} + \mu_2 p_{\mathbf{n}+2} = 0$$

(c) For $(\mathbf{n+1})_2$

$$\frac{\lambda}{2} p_{\mathbf{n}+1} - (\lambda + \mu_1 + \mu_2) p_{(\mathbf{n}+1)} + \mu_1 p_{\mathbf{n}+2} = 0.$$

2.3 The Coefficient Matrix

For later purposes we define the following matrices:

$$\Lambda_0=\begin{pmatrix} -\lambda & \mu_1 & \mu_2 & 0 \\ \frac{\lambda}{2} & -(\lambda+\mu_1) & 0 & \mu_2 \\ \frac{\lambda}{2} & 0 & -(\lambda+\mu_2) & \mu_1 \end{pmatrix}$$

and for $n>0$

$$\Lambda_n=\begin{pmatrix} \lambda & \lambda & -(\lambda+\mu_1+\mu_2) & (\mu_1+\mu_2) & (\mu_1+\mu_2) & 0 \\ 0 & 0 & \frac{\lambda}{2} & -(\lambda+\mu_1+\mu_2) & 0 & \mu_2 \\ 0 & 0 & \frac{\lambda}{2} & 0 & -(\lambda+\mu_1+\mu_2) & \mu_1 \end{pmatrix}, \quad n=1, 2, \ldots.$$

We partition these matrices as

$$\Lambda_{01}=\begin{pmatrix} -\lambda \\ \frac{\lambda}{2} \\ \frac{\lambda}{2} \end{pmatrix}, \qquad \Lambda_{02}=\begin{pmatrix} \mu_1 & \mu_2 & 0 \\ -(\lambda+\mu_1) & 0 & \mu_2 \\ 0 & -(\lambda+\mu_2) & \mu_1 \end{pmatrix},$$

$$\Lambda_{n1}=\begin{pmatrix} \lambda & \lambda & -(\lambda+\mu_1+\mu_2) \\ 0 & 0 & \frac{\lambda}{2} \\ 0 & 0 & \frac{\lambda}{2} \end{pmatrix}, \text{ and } \quad \Lambda_{n2}=\begin{pmatrix} (\mu_1+\mu_2) & (\mu_1+\mu_2) & 0 \\ -(\lambda+\mu_1+\mu_2) & 0 & \mu_2 \\ 0 & -(\lambda+\mu_1+\mu_2) & \mu_1 \end{pmatrix}$$

Then the coefficient matrix for the general "steady state" equations $\Lambda P=0$ can be written as (for $r=2$)

$$\Lambda=\begin{pmatrix} \Lambda_{01} & \Lambda_{02} & 0 & 0 & 0 & 0 \ldots \\ 0 & \Lambda_{11} & \Lambda_{12} & 0 & 0 & 0 \ldots \\ 0 & 0 & \Lambda_{21} & \Lambda_{22} & 0 & 0 \ldots \\ 0 & 0 & 0 & \Lambda_{31} & \Lambda_{32} & 0 \ldots \end{pmatrix}. \tag{2.1}$$

The matrices Λ_{n1}, Λ_{n2}, $n>0$ are independent of n and hence the partitioned matrix is simply a bi-diagonal matrix with $\Lambda_{j1}=\Lambda_{i1}$ and $\Lambda_{j2}=\Lambda_{i2}$ for $i, j=1, 2, \ldots$. The important consideration, however, is that this partitioned form of the matrix depends only on the existence of the instantaneous jockeying rules of Section 1.2 only. Except for the size of the submatrices, the number of servers has no effect on the structure of Λ. Thus, the fact that we have chosen to carry the structure through for $r=2$ is irrelevant to the construction of the partitioned matrix above and to the solution below. All results are valid for $r\geq 2$. Our choice of $r=2$ was pedagogical only.

2.4 The State Probability Vectors

For $r=2$, let

$$P_{01}=p_0, \text{ a scalar},$$

$$P_{02}^T=(p_{0_1}, p_{0_2}, p_1),$$

P_{02} is a column vector of length 3. In general, P_{0r} is of length 2^r-1. Further, for all n, let

$P_{(n+1)1}^T=(p_{1_{n_1}}, p_{n_2}, p_n^n)$, $n=0, 1, 2, \ldots$. In general, $P_{(n+1)1}$ is a column vector of length $(r+1)$.

$P_{(n+1)2}^T=(p_{(n+1)_1}, p_{(n+1)_2}, p_{n+1})$, $n=0, 1, 2, \ldots$. In general, $P_{(n+1)2}$ will be of length 2^r-1.

2.5 The Steady State Equations

The steady state equations can be written using the above partitions as

$$\wedge P=0,$$

where $\wedge$ is defined by (2.1) and P is the column vector whose elements are given in Section 2.4. More importantly, however, one has

$$\wedge_{01}P_{01}+\wedge_{02}P_{02}=0 \tag{2.2}$$

$$\wedge_{n1}P_{(n+1)1}+\wedge_{n2}P_{(n+1)2}=0,\ n=0, 1, 2, \ldots. \tag{2.3}$$

Again, this set of equations does not depend on r. Hence every instantaneous jockeying queue satisfying our assumption of Section 2 (and extensions given in Section 5) satisfy this system of equations.

3.0 THE SOLUTIONS

Using the partitioned form of Section 2.5 it follows directly that

$$P_{02}=-\wedge_{02}^{-1}\wedge_{01}P_{01}, \tag{3.1}$$

for any $r>1$.

Define

$$B=\wedge_{02}^{-1}\wedge_{01}.$$

B is $(2^r-1)\times 1$. Let B_1 be the vector of the first (2^r-r-2) rows and B_2 the remaining $(r+1)$ rows of B. Then (3.1) is equivalent to

$$P_{02}=-\begin{pmatrix}B_1P_{01}\\ B_2P_{01}\end{pmatrix}. \tag{3.2}$$

From (2.3) one has

$$P_{(n+1)2}=-\wedge_{n2}^{-1}\wedge_{n1}P_{(n+1)1}, \qquad n=0, 1, 2, \ldots. \tag{3.3}$$

Equation (3.3) defines the (2^r-1) elements of $P_{(n+1)2}$ in terms of the $(r+1)$ elements of $P_{(n+1)1}$, but the elements of $P_{(n+1)1}$ are known, since they represent the last $(r+1)$ elements of $P_{n,2}$. We make the following definition:

Let: $P^*_{n,2}$ = the last $(r+1)$ rows of $P_{n,2}$. It then follows that

$$P_{(n+1)1} = P^*_{n,2}. \tag{3.4}$$

Using Eqs. (3.1) and (3.3) and iterating, we find

$$P_{02} = -\wedge_{02}^{-1}\wedge_{01}P_{01} \tag{3.5}$$

and

$$P_{12} = -\wedge_{12}^{-1}\wedge_{11}P_{11}.$$

By (3.2) we have

$$P^*_{02} = -B_2P_{01},$$

but by (3.4)

$$P_{11} = P^*_{02} = -B_2P_{01}. \tag{3.6}$$

Thus,

$$P_{12} = \wedge_{12}^{-1}\wedge_{11}B_2P_{01}. \tag{3.7}$$

Since

$$\wedge_{j2}^{-1}\wedge_{j1} = \wedge_{i2}^{-1}\wedge_{i1} \qquad \text{for all } i, j \neq 0, \tag{3.8}$$

we let

$$A = \wedge_{n2}^{-1}\wedge_{n1}.$$

A is a $(2^r-1)x(r+1)$ matrix. Partition A into A_1, A_2 where A_2 is the $(r+1)x(r+1)$ matrix consisting of the last $(r+1)$ columns of A.

We can assemble these terms to find all the unknown probabilities in terms of P_{01} – a scalar. P_{02} is given by (3.2) in terms of known matrices and P_{01}. Also by (3.6) and (3.7) one has the value of P_{11}, P_{12} in terms of the given matrices and P_{01}. Using (3.6), A_1 and A_2, (3.1) and (3.3), one has

$$P_{12} = -\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}P_{11}.$$

which upon using (3.7) gives

$$P_{12} = (-1)^2\begin{pmatrix} A_1B_2P_{01} \\ A_2B_2P_{01} \end{pmatrix}.$$

and by using the definition of $P^*_{n,2}$ we obtain

$$P^*_{12} = A_2B_2P_{01} = P_{21}.$$

And continuing the iteration

$$P_{22} = -AP_{21}$$

$$= (-1)^2\begin{pmatrix} A_1A_2B_2P_{01} \\ A_2^2B_2P_{01} \end{pmatrix}$$

$$P^*_{22} = -A_2^3B_2P_0 = P_{31}$$

$$P_{32} = \begin{pmatrix} A_1A_2^2B_2P_{01} \\ A_2^3B_2P_{01} \end{pmatrix}.$$

In general,

$$P_{(n+1)_2} = (-1)^{n+1}\begin{pmatrix} A_1 A_2^n B_2 P_{01} \\ A_2^{n+1} B_2 P_{01} \end{pmatrix} \text{ for all } n \geq 1.$$

(note: $A_2^0 = I$)

$$P_{n2}^* = (-1)^{n+1} A_2^n B_2 P_{01}. \tag{3.9}$$

Thus, a complete solution for all state probabilities is given by:

$$P_{02} = -BP_{01} \tag{3.10}$$

$$P_{(n+1)_2} = (-1)^{n+1}\begin{pmatrix} A_1 A_2^n B_2 P_{01} \\ A_2^{n+1} B_2 P_{01} \end{pmatrix}, \; n = 0, 1, 2, \ldots \tag{3.11}$$

$$P_{(n+1)_1} = (-1)^n (A_2^n B_2 P_{01}). \tag{3.12}$$

Equations (3.2), (3.10), and (3.11) give us all state probabilities in a closed matrix form in terms of the single scalar P_0. The probabilities given by (3.12) are redundant and are included in other vectors. Hence they do not comprise a part of the set of state probabilities. P_{01} is determined by requiring all terms to sum to 1.

It is useful to note that all probabilities are obtained in terms of A and B and that A requires one inversion, of the matrix $\wedge_{n2}$, while B requires one inversion of the matrix $\wedge_{02}$.

4.0 AN EXAMPLE: THE CASE r=2 (Haight [2], Koenigsberg [3]).*

These equations and their associated matrices have been given in Section 2. Here we note:

$$\wedge_{02}^{-1} = \begin{pmatrix} \dfrac{\lambda+\mu_2}{\mu_1(2\lambda+\mu_1+\mu_2)} & -\dfrac{1}{(2\lambda+\mu_1+\mu_2)} & \dfrac{\mu_2}{\mu_1(2\lambda+\mu_1+\mu_2)} \\ \dfrac{\lambda+\mu_1}{\mu_2(2\lambda+\mu_1+\mu_2)} & \dfrac{\mu_1}{\mu_2(2\lambda+\mu_1+\mu_2)} & -\dfrac{1}{(2\lambda+\mu_1+\mu_2)} \\ \dfrac{(\lambda+\mu_1)(\lambda+\mu_2)}{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)} & \dfrac{\lambda+\mu_2}{\mu_2(2\lambda+\mu_1+\mu_2)} & \dfrac{\lambda+\mu_1}{\mu_1(2\lambda+\mu_1+\mu_2)} \end{pmatrix}$$

and from (3.2)

$$P_{02} = \wedge_{02}^{-1} \wedge_{01} P_0 = \begin{pmatrix} \dfrac{\lambda}{2\mu_1} \\ \dfrac{\lambda}{2\mu_2} \\ \dfrac{\lambda^2}{2\mu_1\mu_2} \end{pmatrix} P_0,$$

*As pointed out by Koenigsberg [3: p 422] the results of this section are identical to those obtained by Gumbel [1] for the maitre d'hotel system with two heterogeneous servers.

or

$$P_{0_1}=\frac{\lambda}{2\mu_1}p_0,$$

$$P_{0_2}=\frac{\lambda}{2\mu_2}p_0,$$

and

$$P_1=\frac{\lambda^2}{2\mu_1\mu_2}p_0.$$

These values agree with those previously given.

Similarly, for the general equations, we find

$$\wedge_{n2}^{-1}=\begin{pmatrix}\frac{\mu_2}{(\mu_1+\mu_2)^2} & -\frac{\mu_1}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)} & \frac{\mu_2}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)}\\ \frac{\mu_1}{(\mu_1+\mu_2)^2} & \frac{\mu_1}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)} & -\frac{\mu_2}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)}\\ \frac{\lambda+\mu_1+\mu_2}{(\mu_1+\mu_2)^2} & \frac{1}{\mu_1+\mu_2} & \frac{1}{\mu_1+\mu_2}\end{pmatrix}$$

and

$$\wedge_{n2}^{-1}\wedge_{n1}=\begin{pmatrix}\frac{\lambda\mu_1}{(\mu_1+\mu_2)^2} & \frac{\lambda\mu_1}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)} & -\frac{2\mu_1(\lambda+\mu_1+\mu_2)^2+\lambda(\mu_2-\mu_1)^2}{2(\mu_1+\mu_2)^2(\lambda+\mu_1+\mu_2)}\\ \frac{\lambda\mu_2}{(\mu_1+\mu_2)^2} & -\frac{\lambda\mu_1}{(\mu_1+\mu_2)(\lambda+\mu_1+\mu_2)} & -\frac{2\mu_2(\lambda+\mu_1+\mu_2)^2+\lambda(\mu_1-\mu_2)^2}{2(\mu_1+\mu_2)^2(\lambda+\mu_1+\mu_2)}\\ \frac{\lambda(\lambda+\mu_1+\mu_2)}{(\mu_1+\mu_2)^2} & \frac{\lambda}{\mu_1+\mu_2} & \frac{(\lambda+\mu_1+\mu_2)^2+\lambda(\mu_1+\mu_2)}{(\mu_1+\mu_2)^2}\end{pmatrix}$$

Notice, for $r=2$ only,

$$\wedge_{n2}^{-1}\wedge_{n1}=A=A_2,$$

i.e., the last $(r+1)$ rows of A comprise the whole matrix.

Thus, the general solution from (3.11) is

$$P_{(n+1)2}=(-1)^{n+1}A^{n+1}B_2P_{01},$$

which, in the form given by Haight [2] and Koenigsberg [3], is:

$$P_n=\frac{\lambda^2\rho^{2(n+1)}}{2\mu_1\mu_2}p_0,\ P_{(n+1)_1}=\frac{\lambda(\lambda+2\mu_1\rho^2)\rho^{2n}}{4\mu_1\mu_2\rho(1+\rho)}p_0,$$

and

$$P_{(n+1)_2}=\frac{\lambda(\lambda+2\mu_2\rho^2)\rho^{2n}}{4\mu_1\mu_2\rho(1+\rho)}p_0,$$

where:

$$\rho=\frac{\lambda}{\mu_1+\mu_2}.$$

5.0 EXTENSIONS*

5.1 Queue Selection Preference†

Suppose that the customer has some preference for one of the possible choices such that the probability of joining an "open" queue is not $1/s$ (see 1.1(d)).

"open" queue is not $1/s$ (see 1.1(d)).

We give the following definition: An open queue is one such that if an arrival joins that queue, it will not give an impossible state. (Note: all queues in the state n are open queues; if all queues are in the state $n+1$ then those queues are also open.)

Let

$$_i\pi_{ij\ldots s} = \text{Prob (joining the } i\text{th queue} \mid i\text{th. } j\text{th, . . ., } s\text{th queues only are open),}$$

$$_i\pi_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise,} \end{cases}$$

and

$_i\pi_{j\ldots s}=0$ if i does not appear in both subscripts.

For $r=2$, the Λ matrices are given by:

$$\Lambda_0 = \begin{pmatrix} -\lambda & \mu_1 & \mu_2 & 0 \\ {}_1\pi_{12}\lambda & -(\lambda+\mu_1) & 0 & \mu_2 \\ {}_2\pi_{12}\lambda & 0 & -(\lambda+\mu_2) & \mu_1 \end{pmatrix}$$

and

$$\Lambda_n = \begin{pmatrix} \lambda & \lambda & -(\lambda+\mu_1+\mu_2) & (\mu_1+\mu_2) & (\mu_1+\mu_2) & 0 \\ 0 & 0 & {}_1\pi_{12}\lambda & -(\lambda+\mu_1+\mu_2) & 0 & \mu_2 \\ 0 & 0 & {}_2\pi_{12}\lambda & 0 & -(\lambda+\mu_1+\mu_2) & \mu_1 \end{pmatrix}$$

These terms do not change the form of the coefficient matrix Λ (it does change some of the coefficients of the state probabilities, specifically those terms that deal with arrivals to the system). The general solution of Section 3 remains valid.

5.2 Balking

We define

$$_0\Pi_{ij\ldots s} = \text{Prob (balking/}i\text{th. } j\text{th, . . ., } s\text{th queues are open),}$$

and $_i\Pi_{ij\ldots s}$ is defined as in Section 5.1, for $i\neq 0$. If we require $_0\Pi_{ij\ldots s}=1$ whenever all queues are of size N then Λ represents the coefficient matrix for the queueing system with finite (rN) capacity. We note that balking probabilities do not directly enter into or change the state equations in any way, except that the π's that do appear no longer add to unity, as in Section 5.1. This merely reflects the fact that the customer no longer joins a queue with probability one. Hence the general solution of Section 3 remains valid.

---

*In addition to the extensions given explicitly here, one can imagine other possible behaviors that modify the transition rates, but retain the basic structure of the Λ matrix. For example, reneging can be incorporated without losing the structure. Such inclusion is obvious and we do not explicitly expose the details.

† Krishnamoorthi [4] has given results for this selection rule for the two server case.

5.3 Customer Jockeying Preference

Another generalization involves the jockeying discipline itself. It will be remembered that, if s possible jockeys were available, then each would occur with probability $1/s$. (Jockeying Rule 1.2(b)) We note that such a situation could only occur (two or more available jockeys) for $r \geq 3$. For example, for $r=3$ suppose the state was $\mathbf{n}_{12}$ and a service occurred in the 3rd queue, giving the instantaneous state $(n+1, n+1, n-1)$. Two possible jockeys could occur, from either the 1st or 2nd queue (but not both). Initially, we defined each of these to happen with probability 1/2. But now let us suppose that there is a probability distribution on these jockeying choices. In effect, we are now allowing jockeying preferences; in turn, this allows us to consider the distance the jockeyer must travel; it is now possible to take explicitly into account the fact that a person in an adjacent queue is more likely to jockey than a person from a distant queue.

DEFINITION: An eligible queue i is one in which the difference $n_i - n_k \geq 2$; in other words, the ith queue contains a customer who may jockey. Let us define the following:

$${}_{jk}\alpha_{ij\ldots c} = \text{Prob (jockeying from queue } j \text{ to } k/\text{the queues } i, j, \ldots, c \text{ only are eligible to jockey).}$$

Furthermore, let

$${}_{jk}\alpha_{ij\ldots c} = \begin{cases} 0 & \text{if } j \text{ does not appear in both subscripts} \\ 1 & \text{if } j=k \\ 0 & \text{if } k \text{ appears in both subscripts} \\ {}_{jk}\alpha_{ij\ldots c} & \text{otherwise.} \end{cases}$$

The α's only affect the equations for $\mathbf{n} \geq 1$ since no jockeying occurs under the initial conditions since all arrivals immediately enter service. Again the structure of the problem given in section 3 is unaffected by this change and the solution given there remains valid.

5.4 Dependence on n, the Number in the Queue

Suppose that Λ_n becomes a function of n. In other words, the arrival rates, the service rates, the queue preference probabilities, or the customer jockeying probabilities now become dependent on the number of people in a queue. By a development which parallels that of Section 3 it can be shown that the solution is given by

$$P_{n2} = \begin{pmatrix} A_1(n)A_2(1)A_2(2) \ldots\ldots A_2(n-1)B_2P_0 \\ A_2(1)A_2(2) \ldots\ldots\ldots A_2(n)B_2P_0 \end{pmatrix}.$$

This follows by dropping the condition (3.8) and letting

$$\Lambda_{n2}^{-1}\Lambda_{n1} = A(n).$$

We then partition $A(n)$ into

$$A(n) = \begin{pmatrix} A_1(n) \\ A_2(n) \end{pmatrix}.$$

By iteration, we find the solution to be as given above.

6.0 CONCLUSIONS

We have given a solution technique which seems to be powerful for a large class of jockeying problems. A simple matrix equation has given us a closed form solution for any number of servers. Simple extensions of the method allowed us to include the problems of customer queue selection preference, jockeying preference, dependence on the number in the queue, and balking, where the balking case included the finite capacity queue as a special case.

The driving force in the system, in all of its forms, is the instantaneous jockeying principle. This principle allows us to cast the steady state equations in their readily solvable form. This solution requires a customer to jockey if it is possible. The refinements presented in Section 5 retain the special structure of Λ and hence do not present important modifications to those solutions given by Eqs. (3.10) and (3.11).

REFERENCES

[1] Gumbel, H., "Waiting Lines with Heterogeneous Servers," Operations Research, **8,** 504 (1960).

[2] Haight, F. A., "Two Queues in Parallel," Biometrika, **45,** 401 (1958).

[3] Koenigsberg, E., "On Jockeying in Queues," Management Science, **12,** 412 (1966).

[4] Krishnamoorthi, B., "On a Poisson Queue with Two Heterogeneous Servers," Operations Research, **11,** 321 (1963).

THE DISTRIBUTION OF THE PRODUCT OF TWO NONCENTRAL BETA VARIATES

Henrich John Malik

University of Guelph
Guelph, Ontario

ABSTRACT

In this paper the exact distribution of the product of two noncentral beta variates is derived using Mellin integral transform. The density function of the product is represented as a mixture of Beta distributions and the distribution function as a mixture of Incomplete Beta Functions.

1. INTRODUCTION

Mellin transform is a powerful analytical tool in studying the distribution of products and quotients of independent random variables. The operational advantages of Mellin transforms in problems of this type have been discussed by Epstein [3]. Following Epstein many authors applied the Mellin transform in a number of papers on the distribution of products and quotients of random variables: a detailed bibliography can be found in Springer and Thompson [8]. Examples of engineering applications involving products and quotients of random variables can be found in Donahue [2]. The practical usefulness of the results described above is limited by the fact that all the corresponding distributions have infinite ranges while in many physical applications the mathematical models often have finite characteristics.

The situation involving product of independent Beta variates arises in many applications, for instance, in system reliability. If it is assumed that the system consists of a number of subsystems and the initial reliability estimated from each subsystem, R_i, suggests a Beta density, then the total reliability, $R=R_1R_2 \ldots R_N$, is a random variable, and it is important to know the distribution of this product. This paper gives the exact distribution of the product of two noncentral Beta variates.

2. THE DISTRIBUTION OF THE PRODUCT OF TWO NONCENTRAL BETA VARIATES.

Let y_1 and y_2 be two independent random variables distributed according to the noncentral beta density function [4] with parameters p_1, q_1, λ_1 and p_2, q_2, λ_2, respectively. Thus the density function of y_j is

$$(1) \qquad q(y_j;\, p_j,\, q_j,\, \lambda_j)=\sum_{i=0}^{\infty}\frac{\Gamma\left(\frac{2i+p_j+q_j}{2}\right)\lambda_j^i}{\Gamma\left(\frac{q_j}{2}\right)\Gamma\left(\frac{2i+p_j}{2}\right)i!}e^{-\lambda_j}y^{1/2(2i+p_j-2)}(1-y_j)^{1/2(q_j-2)} \quad j=1,\,2,\; 0\leq y_j\leq 1.$$

We want to find the probability density function of the variate $u=y_1y_2$, by the use of Mellin transforms.

The Mellin transform $f(s)$, corresponding to a function $f(x)$ defined only for $x \geq 0$, is

(2) $$f(s)=\int_0^\infty x^{s-1}f(x)dx.$$

The inverse Mellin transform enabling one to go from the transform $f(s)$ to the function $f(x)$, is

(3) $$f(x)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}x^{-s}f(s)dx.$$

Therefore the Mellin transform of the density function of y_j is

(4) $$f_j(s)=\sum_{i=0}^{\infty}\frac{\Gamma\left(\frac{2i+p_j+q_j}{2}\right)\lambda_j^i e^{-\lambda_j}}{\Gamma\left(\frac{q_j}{2}\right)\Gamma\left(\frac{2i+p_j}{2}\right)i!}\int_0^1 y_j^{1/2(2i+p_j-2)+s-1}(1-y_j)^{1/2(q_j-2)}dy_j.$$

Term by term integration is justified since the series can be shown to converge uniformly. Therefore, we have

(5) $$f_j(s)=\sum_{i=0}^{\infty}\frac{\Gamma\left(\frac{2i+p_j+q_j}{2}\right)\lambda_j^i e^{-\lambda_j}\Gamma\left(\frac{2i+p_j+2s-2}{2}\right)}{\Gamma\left(\frac{2i+p_j}{2}\right)\Gamma\left(\frac{2i+p_j+q_j+2s-2}{2}\right)i!}.$$

If we take the limit as $\lambda_j\to 0$, the result is

$$\lim_{\lambda_j\to 0}f_j(s)=\frac{\Gamma\left(\frac{p_j+q_j}{2}\right)\Gamma\left(\frac{p_j+2s-2}{2}\right)}{\Gamma\left(\frac{p_j}{2}\right)\Gamma\left(\frac{p_j+q_j+2s-2}{2}\right)},$$

which is the Mellin transform of the central beta distribution with parameters $p_j/2$ and $q_j/2$.

The Mellin transform of the density function of the product of two independent random variables is the product of the Mellin transforms of the density functions of the individual variables [2]; therefore, the Mellin transform of the density function of $u=y_1y_2$ is

$$f(s)=f_1(s)f_2(s)$$

(6)
$$=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{\infty}\frac{\lambda_1^i\Gamma\left(\frac{2i+p_1+q_1}{2}\right)\Gamma\left(\frac{2i+p_1+2s-2}{2}\right)}{\Gamma\left(\frac{2i+p_1}{2}\right)\Gamma\left(\frac{2i+p_1+q_1+2s-2}{2}\right)i!}\cdot\sum_{k=0}^{\infty}\frac{\lambda_2^k\Gamma\left(\frac{2k+p_2+q_2}{2}\right)\Gamma\left(\frac{2k+p_2+2s-2}{2}\right)}{\Gamma\left(\frac{2k+p_2}{2}\right)\Gamma\left(\frac{2k+p_2+q_2+2s-2}{2}\right)k!}$$

$$=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{\infty}\sum_{k=0}^{i}\frac{\lambda_1^k\Gamma\left(\frac{2k+p_1+q_1}{2}\right)\Gamma\left(\frac{2k+p_1+2s-2}{2}\right)}{\Gamma\left(\frac{2k+p_1}{2}\right)\Gamma\left(\frac{2k+p_1+q_1+2s-2}{2}\right)k!}$$

$$\cdot\frac{\lambda_2^{i-k}\Gamma\left(\frac{2i-2k+p_2+q_2}{2}\right)\Gamma\left(\frac{2i-2k+p_2+2s-2}{2}\right)}{\Gamma\left(\frac{2i-2k+p_2}{2}\right)\Gamma\left(\frac{2i-2k+p_2+q_2+2s-2}{2}\right)(i-k)!}.$$

where the sum over m comes from the hypergeometric function.

The distribution function of u is given by

$$F(u)=\int_0^u f(t)dt=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{\infty}\sum_{k=0}^{i}\sum_{m=0}^{\infty}$$

$$\frac{\lambda_1^{i}\lambda_2^{i-k}\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}+k\right)\Gamma\left(\frac{q_2}{2}+m\right)\Gamma\left(2k-i+\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2}+m\right)}{\Gamma\left(2k-i+\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2}\right)\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}+\frac{q_2}{2}+k+m\right)\beta\left(\frac{p_2}{2}+i-k,\frac{q_2}{2}\right)k!(i-k)!m!}$$

$$\cdot I_u\left(\frac{p_1}{2}+k,\frac{q_1}{2}+\frac{q_2}{2}+m\right), \tag{11}$$

where

$$I_u(a,b)=\frac{1}{\beta(a,b)}\int_0^u t^{a-1}(1-t)^{b-1}dt$$

is the Incomplete Beta Function.

If we set $\lambda_1=\lambda_2=0$ in (9), the density function of two central beta variates is

$$h(u)=\frac{\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}\right)\Gamma\left(\frac{p_2}{2}+\frac{q_2}{2}\right)u^{\frac{p_1}{2}-1}(1-u)^{\frac{p_1}{2}+\frac{q_2}{2}-1}}{\Gamma\left(\frac{p_1}{2}\right)\Gamma\left(\frac{p_2}{2}\right)\Gamma\left(\frac{q_1}{2}+\frac{q_2}{2}\right)}\quad F\left(\frac{q_2}{2},\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2};\frac{q_1}{2}+\frac{q_2}{2};1-u\right).$$

ACKNOWLEDGMENT

The author is thankful to the referee for his helpful comments.

REFERENCES

[1] Consul, P. C., "On Some Inverse Mellin Transforms," Bull. Class. des Sc. **52,** 547–561 (1966).

[2] Donahue, J. D., *Products and Quotients of Random Variables and Their Applications* (Office of Aerospace Research, USAF, 1964).

[3] Epstein, B., "Some Applications of the Mellin Transform in Statistics," Ann. Math. Statist. **19,** 370–379 (1948).

[4] Graybill, F. A., *An Introduction to Linear Statistical Models* (McGraw-Hill Book Company, Inc., New York, 1961), Vol. 1.

[5] Malik, H. J., "Exact Distribution of the Quotient of Independent Generalized Gamma Random Variates," Can. Math. Bull., **10,** 463–465 (1967).

[6] Malik, H. J., "Exact Distribution of the Product of Two Generalized Gamma Variates," Ann. Math. Statist. **39,** 1751–1752 (1968).

[7] Rao, C. R., *Introduction to Statistical Inference and its Applications* (John Wiley & Sons, New York, 1965).

[8] Springer, M. D. and W. E. Thompson, "The Distribution of Product of Independent Random Variables," Siam J. Appl. Math. **14,** 511–526 (1966).

[9] Wells, W. T., R. L. Anderson, and J. W. Cell, "The Distribution of the Product of Two Central or Noncentral Chi-Square Variates," Ann. Math. Statist. **33,** 1016–1020 (1962).

where the sum over m comes from the hypergeometric function.

The distribution function of u is given by

$$F(u)=\int_0^u f(t)\,dt=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{\infty}\sum_{k=0}^{i}\sum_{m=0}^{\infty}$$

$$\frac{\lambda_1^{i-k}\lambda_2^{k}\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}+k\right)\Gamma\left(\frac{q_2}{2}+m\right)\Gamma\left(2k-i+\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2}+m\right)}{\Gamma\left(2k-i+\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2}\right)\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}+\frac{q_2}{2}+k+m\right)\beta\left(\frac{p_2}{2}+i-k,\frac{q_2}{2}\right)k!(i-k)!m!}$$

$$\cdot I_u\left(\frac{p_1}{2}+k,\frac{q_1}{2}+\frac{q_2}{2}+m\right), \tag{11}$$

where

$$I_u(a,b)=\frac{1}{\beta(a,b)}\int_0^u t^{a-1}(1-t)^{b-1}dt$$

is the Incomplete Beta Function.

If we set $\lambda_1=\lambda_2=0$ in (9), the density function of two central beta variates is

$$h(u)=\frac{\Gamma\left(\frac{p_1}{2}+\frac{q_1}{2}\right)\Gamma\left(\frac{p_2}{2}+\frac{q_2}{2}\right)u^{\frac{p_1}{2}-1}(1-u)^{\frac{p_1}{2}+\frac{q_2}{2}-1}}{\Gamma\left(\frac{p_1}{2}\right)\Gamma\left(\frac{p_2}{2}\right)\Gamma\left(\frac{q_1}{2}+\frac{q_2}{2}\right)}\quad F\left(\frac{q_2}{2},\frac{p_1}{2}-\frac{p_2}{2}+\frac{q_1}{2};\frac{q_1}{2}+\frac{q_2}{2};1-u\right).$$

ACKNOWLEDGMENT

The author is thankful to the referee for his helpful comments.

REFERENCES

[1] Consul, P. C., "On Some Inverse Mellin Transforms," Bull. Class. des Sc. **52**, 547–561 (1966).

[2] Donahue, J. D., *Products and Quotients of Random Variables and Their Applications* (Office of Aerospace Research, USAF, 1964).

[3] Epstein, B., "Some Applications of the Mellin Transform in Statistics," Ann. Math. Statist. **19**, 370–379 (1948).

[4] Graybill, F. A., *An Introduction to Linear Statistical Models* (McGraw-Hill Book Company, Inc., New York, 1961), Vol. 1.

[5] Malik, H. J., "Exact Distribution of the Quotient of Independent Generalized Gamma Random Variates," Can. Math. Bull., **10**, 463–465 (1967).

[6] Malik, H. J., "Exact Distribution of the Product of Two Generalized Gamma Variates," Ann. Math. Statist. **39**, 1751–1752 (1968).

[7] Rao, C. R., *Introduction to Statistical Inference and its Applications* (John Wiley & Sons, New York, 1965).

[8] Springer, M. D. and W. E. Thompson, "The Distribution of Product of Independent Random Variables," Siam J. Appl. Math. **14**, 511–526 (1966).

[9] Wells, W. T., R. L. Anderson, and J. W. Cell, "The Distribution of the Product of Two Central or Noncentral Chi-Square Variates," Ann. Math. Statist. **33**, 1016–1020 (1962).

OPTIMUM ALLOCATION OF QUANTILES IN DISJOINT INTERVALS FOR THE BLUES OF THE PARAMETERS OF EXPONENTIAL DISTRIBUTION WHEN THE SAMPLE IS CENSORED IN THE MIDDLE

A. K. Md. Ehsanes Saleh*

Carleton University, Ottawa

and

M. Ahsanullah†

Food and Drug Directorate, Ottawa

1. INTRODUCTION AND SUMMARY

In the theory of estimation it is well known that when all the observations in a sample are available, it is sometimes possible to obtain estimators that are the most efficient linear combinations of a given number of order statistics. In many practical situations we encounter censored samples, that is, samples where values of some of the observations are not available. Singly and doubly censored samples occur when the extreme observations are not available and middle censored samples occur when observations are missing from the middle of an ordered sample. Censoring in the middle of a sample may occur due to measurement restrictions, time, economy or failure of the measuring instrument to record observations or due to off-shifts or week-end interruptions in the course of an experiment. As mentioned in Sarhan and Greenberg [8] in the space telemetry, where signals are supposed to be sent at regular intervals we may expect a few of these signals to be missing during journey and at the end of communication.

In this paper we shall consider the problem of best linear unbiased estimation (BLUE) of the parameters of the exponential distribution based on a fixed number k (less than the number of available observations) selected order statistics when the sample is censored in the middle. The study is based on the asymptotic theory of quantiles and under type II censoring scheme. The optimal allocation of the k quantiles in the two disjoint intervals along with the optimum spacings of the quantiles have been determined. The estimates and their efficiencies may easily be calculated based on the table of coefficients and efficiencies presented at the end of this paper, in Table 1, for various proportions of censoring.

The problem of choice of optimal k quantiles in uncensored and singly and doubly censored samples have been dealt with by Kulldorff, [1, 2] Ogawa, [3] Saleh and Ali, [4] Saleh, [5, 6] and Sarhan and Greenberg [7, 8]. The present problem is an extension to censoring in the middle posing a new problem of optimum allocation of k quantiles in the two disjoint intervals due to censoring in the middle.

*Research supported by the National Research Council of Canada. This work has been completed while the author was a fellow at the Summer Research Institute, McGill University, 1969.

†On leave from Institute of Statistical Research and Training, Dacca University, Dacca, Pakistan.

2. ESTIMATION OF THE PARAMETERS

Suppose we are sampling from the exponential distribution

$$F(x)=1-\exp\left(-\frac{x-\mu}{\sigma}\right),\ x\geq\mu,\ \sigma>0, \tag{2.1}$$

where μ and σ are the parameters of the distribution. The sample size, n, is assumed to be large; le the interval (0, 1) l e sub-divided into three intervals: $I_1=(0,\ \alpha]$, $I_2=(\alpha,\ \beta)$, and $I_3=[\beta,\ 1)$ with $0<\alpha<\beta<1$. Define $p_0=0$ and $p_3=1$ so that $p_0=0<p_1=\alpha<\beta=p_2<p_3=1$. Under type II censoring scheme in the middle, we only retain α and $1-\beta$ proportion of samples from the two extreme intervals so that the proportion of censoring is $\beta-\alpha$. Thus the ranks of all the uncensored observations lie in the intervals $[1,\ n_1]$ and $[n_2,\ n]$, respectively, where $n_1=[n\alpha]+1$ and $n_2=[n\beta]+1$ and $[\ \ ]$ is the Euler's notation for the largest integer c ntained in $[\ \ ]$. In this section, we shall obtain the BLUES of the parameters based on k arbitrary quantiles whose ranks are available from the two disjoint integer sets $[1,\ n_1]$ and $[n_2,\ n]$, respectively.

Let the ordered observations in a sample of size n be $x_{(1)}<x_{(2)}\ .\ .\ .\ <x_{(n)}$ and consider the k sample quantiles $x_{(n_{11})}<x_{(n_{12})}<\ .\ .\ .\ <x_{(n_{1k_1})}<x_{(n_{21})}<\ .\ .\ .\ <x_{(n_{2k_2})}$, where the ranks n_{ij} are given by

$$n_{1j}=[np_{1j}]+1 \qquad j=1,\ 2,\ .\ .\ .\ k_1 \tag{2.2a}$$

and

$$n_{2j}=[np_{2j}]+1 \qquad j=1,\ 2,\ .\ .\ .\ k_2, \tag{2.2b}$$

and the spacings $p_{ij}(i=1,\ 2, j=1,\ 2,\ .\ .\ .\ k_i)$ satisfy the inequality

$$0<p_{11}<\ .\ .\ .\ <p_{1k_1}<p_{21}<\ .\ .\ .\ <p_{2k_2}<1; \tag{2.3}$$

also

$$0<p_{1j}\leq\alpha \text{ and } \beta\leq p_{2j}<1 \text{ for all } j.$$

Now if the spacings are redesignated as

$$\lambda_1<\ .\ .\ .\ <\lambda_k,\ k=k_1+k_2,$$

then the expressions for the BLUES and their variances and covariance and the generalized variance will coincide with the expression in (2.7a) through (2.8) of Saleh [5] with necessary restriction due to censoring in the middle.

The symbols $u_{ij}=\ln(1-p_{ij})^{-1}$, $i=1,\ 2, j=1,\ 2,\ .\ .\ .\ k_i$ explain the connections of the expressions which are:

$$\hat{\sigma}=\sum_{j=1}^{k_1}b_{1j}x_{(n_{1j})}+\sum_{j=1}^{k_2}b_{2j}x_{(n_{2j})}, \tag{2.4}$$

(2.5) $$\hat{\mu}=x_{(n_{11})}-\hat{\sigma}u_{11}.$$

(2.6a) $$b_{11}=-\frac{u_{12}-u_{11}}{(e^{u_{12}}-e^{u_{11}})L},$$

(2.6b) $$b_{ij}=L^{-1}\left\{\frac{u_{ij}-u_{1j-1}}{e^{u_{1j}}-e^{u_{1j-1}}}-\frac{u_{1j+1}-u_{1j}}{e^{u_{1j+1}}-e^{u_{1i}}}\right\}$$

where

$$j=2, 3, \ldots k_1 \text{ and } u_{1k_1+1}=u_{21},\ u_{10}=0,$$

(2.6c) $$b_{2j}=L^{-1}\left\{\frac{u_{2j}-u_{2j-1}}{e^{u_{2j}}-e^{u_{2j-1}}}-\frac{u_{2j+1}-u_{2j}}{e^{u_{2j+1}}-e^{u_{2j}}}\right\}$$

where

$$j=1, 2, \ldots, k_2-1,\ u_{20}=u_{1k_1},\ u_{2k_2+1}=0,$$

(2.6d) $$b_{2k_2}=L^{-1}\left\{\frac{u_{2k_2}-u_{2k_2-1}}{e^{u_{2k_2}}-e^{u_{2k_2-1}}}\right\},$$

and

(2.6e) $$L=\sum_{j=1}^{k_1-1}\frac{(u_{1j+1}-u_{1j})^2}{e^{u_{1j+1}}-e^{u_{1j}}}+\sum_{j=0}^{k_2-1}\frac{(u_{2j+1}-u_{2j})^2}{e^{u_{2j+1}}-e^{u_{2j}}}+\frac{(u_{21}-u_{1k_1})^2}{e^{u_{21}}-e^{u_{1k_1}}}$$

The variances and covariance of the estimates are

(2.7a) $$V(\hat{\mu})=\frac{\sigma^2}{n}\{L^{-1}u_{11}^2+(e^{u_{11}}-1)\},$$

(2.7b) $$V(\hat{\sigma})=\frac{\sigma^2}{n}L^{-1},$$

and

(2.7c) $$cov\,(\hat{\mu},\hat{\sigma})=\frac{\sigma^2}{n}(u_{11}L^{-1}).$$

The generalized variance of the estimate is

(2.7d) $$\Lambda=\frac{\sigma^4}{n^2}(e^{u_{11}}-1)L^{-1}.$$

When $\mu=0$, the estimate σ based on the k quantiles is

(2.8) $$\hat{\sigma}=\sum_{j=1}^{k_1}b_{1j}x_{(n_{1j})}+\sum_{j=2}^{k_2}b_{2j}x_{(n_{2j})},$$

where

(2.9a) $$b_{1j}=Q_k^{-1}\left\{\frac{u_{1j}-u_{1j-1}}{e^{u_{1j}}-e^{u_{1j-1}}}-\frac{u_{1j+1}-u_{1j}}{e^{u_{1j+1}}-e^{u_{1j}}}\right\}$$

$j=1, 2, \ldots k_1$ with $u_{1k_1+1}=u_{21}$,

(2.9b) $$b_{2j}=Q_k^{-1}\left\{\frac{u_{2j}-u_{2j-1}}{e^{u_{2j}}-e^{u_{2j-1}}}-\frac{u_{2j+1}-u_{2j}}{e^{u_{2j+1}}-e^{u_{2j}}}\right\}$$

$j=1, 2, \ldots k_2-1$ with $u_{20}=u_{1k_1}$, $u_{2k_2+1}=0$,

(2.9c) $$b_{2k_2}=Q_k^{-1}\left\{\frac{u_{2k_2}-u_{2k_2-1}}{e^{u_{2k_2}}-e^{u_{2k_2-1}}}\right\}$$

and

(2.10) $$Q_k=\sum_{j=0}^{k_1-1}\frac{(u_{1j+1}-u_{1j})^2}{e^{u_{1j+1}}-e^{u_{1j}}}+\sum_{j=0}^{k_2-1}\frac{(u_{2j+1}-u_{2j})^2}{e^{u_{2j+1}}-e^{u_{2j}}}+\frac{(u_{21}-u_{1k_1})^2}{e^{u_{21}}-e^{u_{1k_1}}}.$$

The variance of the estimate is given by

(2.11) $$V(\hat{\sigma})=\frac{\sigma^2}{nQ_k}.$$

We note that in all the above expressions the restrictions on the u's are

(2.12) $$\left.\begin{aligned}&0<u_{11}<\ldots<u_{1k_1}\leq\ln(1-\alpha)^{-1}\\&\ln(1-\beta)^{-1}\leq u_{21}<\ldots<u_{2k_2}<+\infty\end{aligned}\right\}.$$

3. OPTIMUM QUANTILES FOR ESTIMATION OF PARAMETERS

In order to determine the optimum k quantiles for the BLUES of μ and σ simultaneously, we have to minimize the expression for Λ, the generalized variance of the estimates. Equivalently, we maximize

$$(e^{u_{11}}-1)^{-1}L$$

for variations of $u_{11}, u_{12}, u_{1k_1}, \ldots u_{2k_2}$, with the restrictions (2.12) on the u's and for all combinations of k_1 and k_2, such that $k=k_1+k_2$ (fixed). When $\mu=0$ and σ to be estimated, we maximize Q_k as in (2.10) accordingly.

For the two-parameter problem, we observe that $(e^{u_{11}}-1)^{-1}L$ as a function of u_{11} is monotonically decreasing (Saleh [5]) and the maximum is attained at

$$u_{11}^*=\ln\left\{1-\frac{1}{n+1/2}\right\}^{-1}.$$

Thus the optimum spacing is $p_{11}^*=\dfrac{1}{n+1/2}$, and the optimum rank of the quantile is $n_{11}^*=1$. To determine the remaining $k-1$ quantiles, we maximize $(e^{u_{11}}-1)^{-1}L$ with respect to $u_{12}, u_{13}, u_{1k_1}, u_{21}, \ldots u_{2k_2}$ keeping $u_{11}^\circ=\ln\left\{1-\dfrac{1}{n+1/2}\right\}^{-1}$ fixed and for all combinations of k_1 and k_2, such that, $k_1+k_2=k$ (fixed). Thus we use the following transformations

(3.1) $$\begin{aligned}t_{1j-1}&=u_{1j}-u_{11}^* \qquad j=2, \ldots k_1\\ t_{2j}&=u_{2j}-u_{11}^* \qquad j=1, 2, \ldots k_2.\end{aligned}$$

Then, $(e^{u_{11}}-1)^{-1}L$ reduces to $\frac{n-1/2}{n+1/2}Q_{k-1}$, where

$$Q_{k-1}=\sum_{j=0}^{k_1-2}\frac{(t_{1j+1}-t_{1j})^2}{e^{t_{1j+1}}-e^{t_{1j}}}+\sum_{j=0}^{k_2-1}\frac{(t_{2j+1}-t_{2j})^2}{e^{t_{2j+1}}-e^{t_{2j}}}+\frac{(t_{21}-t_{1k_1-1})^2}{e^{t_{21}}-e^{t_{1k_1-1}}}, \tag{3.2}$$

where

$t_{11}, t_{12}, \ldots, t_{1k_1-1}, t_{21}, \ldots, t_{2k_2}$ satisfy the inequalities

$$0<t_{11}<\ldots<t_{1k_1-1}\leq \ln\left[\frac{n-1/2}{(n+1/2)(1-\alpha)}\right]$$

$$\ln\left[\frac{n-1/2}{(n+1/2)(1-\beta)}\right]\leq t_{21}<\ldots t_{2k_2}<+\infty. \tag{3.3}$$

Thus, the problem of determining the optimum quantiles reduces to choosing the corresponding spacings $\lambda_{11}^{\circ}, \lambda_{12}^{\circ}, \ldots, \lambda_{1k_1-1}^{\circ}, \lambda_{21}^{\circ}, \ldots, \lambda_{2k_2}^{\circ}$, which maximizes Q_{k-1} for variations of $t_{11}, \ldots, t_{1k_1-1}, t_{21}, \ldots, t_{2k_2}$ satisfying (3.3) and for all combinations of k_1 and k_2, such that, $k=k_1+k_2$ (fixed). Therefore we should solve the system of equations

$$\begin{aligned}&\frac{\partial Q_{k-1}}{\partial t_{1j}}=0, \qquad j=1, 2, \ldots k_1-1\\ &\frac{\partial Q_{k-1}}{\partial t_{2j}}=0, \qquad j=1, 2, \ldots k_2,\end{aligned} \tag{3.4}$$

for all combinations of k_1 and k_2, such that $k=k_1+k_2$ subject to the restrictions (3.3) on the t's. Let $t_{1j}^*(j=1, 2, \ldots k_1^*-1)$ and $t_{2j}^*(j=1, 2, \ldots k_2^*)$ be the optimum quantiles which provide maximum of Q_{k-1} among all combinations of integers k_1 and k_2, such that $k_1+k_2=k$. Then, the set of spacings $\lambda_{1j}^*(j=1, \ldots k_1-1)$ and $\lambda_{2j}^*(j=1, 2, \ldots, k_2)$ are determined by the relations

$$\begin{aligned}&t_{1j}^*=\ln(1-\lambda_{1j}^*)^{-1}, \qquad j=1, \ldots k_1-1\\ &t_{2j}^*=\ln(1-\lambda_{2j}^*)^{-1}, \qquad j=1, 2, \ldots k_2.\end{aligned} \tag{3.5}$$

The optimum choice of spacings for the estimation of (μ, σ) are obtained entirely by the relations

$$\left.\begin{aligned}p_{1j+1}^*&=\frac{2+(2n-1)\lambda_{1j}^*}{2n+1}; \qquad j=1, 2, \ldots k_1-1\\ p_{2j}^*&=\frac{2+(2n-1)\lambda_{2j}^*}{2n+1}; \qquad j=1, 2, \ldots k_2\end{aligned}\right\}. \tag{3.6}$$

The optimum ranks of the quantiles selected are given by

$$\left.\begin{aligned} n_{11}^* &= 1 \\ n_{1j}^* &= [np_{1j}^*]+1, \quad j=2, \ldots k_1 \\ n_{2j}^* &= [np_{2j}^*]+1, \quad j=1, \ldots k_2 \end{aligned}\right\}. \tag{3.7}$$

The asymptotic BLUES of μ and σ based on the optimum quantiles are

$$\hat{\mu}=x_{(1)}-\hat{\sigma}\ln\left(\frac{2n+1}{2n-1}\right),$$

$$\hat{\sigma}=b_{11}^* x_{(1)}+\sum_{j=2}^{k_1} b_{1j}^* x_{(n_{1j}^*)}+\sum_{j=1}^{k_2} b_{2j}^* x_{(n_{2j}^*)},$$

where

$$b_{11}^*=-\left[\sum_{j=2}^{k_1} b_{1j}^*+\sum_{j=1}^{k_2} b_{2j}^*\right],$$

where $b_{12}^*, \ldots b_{1k_1}^*$ and $b_{21}^*, \ldots b_{2k_2}^*$ may be determined from Table 1. The asymptotic joint efficiency (JAE) and the asymptotic relative efficiencies (ARE) compared to the best linear estimates using all observations in the censored sample (see Sarhan and Greenberg [7]) are given by

$$\text{JAE}\ (\hat{\mu},\hat{\sigma})=\frac{2n-1}{2n}\,\frac{Q_{k-1}^*(\beta-\alpha)}{(\beta-\alpha)(1+\alpha-\beta)+(1-\alpha)(1-\beta)\{\ln(1-\alpha)^{-1}-\ln(1-\beta)^{-1}\}^2},$$

(3.8a)

$$\text{ARE}\ (\hat{\sigma})=\frac{Q_{k-1}^*(\beta-\alpha)}{(\beta-\alpha)(1+\alpha-\beta)+(1-\alpha)(1-\beta)\{\ln(1-\alpha)^{-1}-\ln(1-\beta)^{-1}\}^2}, \tag{3.8b}$$

and

$$\text{ARE}\ (\hat{\mu})=\frac{1}{n}\,\frac{(2n-1)Q_{k-1}^*}{\left[(2n-1)\ln^2\dfrac{2n+1}{2n-1}+2Q_{k-1}^*\right]} \tag{3.8c}$$

$$\left[1+\frac{1}{n}\,\frac{\beta-\alpha}{(\beta-\alpha)(1+\alpha-\beta)+(1-\alpha)(1-\beta)\{\ln(1-\alpha)^{-1}-\ln(1-\beta)^{-1}\}^2}\right],$$

where Q_{k-1}^* is the maximum value of Q_{k-1} defined at (3.2). Thus, once Q_{k-1}^* is known, the efficiencies can easily be computed. We must note that the above asymptotic efficiencies have been computed using the large sample approximation of the generalized variance and the variances of the estimates using all the uncensored observations presented in Sarhan and Greenberg [7] (pp. 357–360). The following example has been presented with finite sample size to illustrate the estimation procedure.

EXAMPLE – Simultaneous estimation of μ and σ: Assume $n=62$, $k=8$, $\alpha=0.4096$, $\beta=0.7048$, and $\beta-\alpha=0.2952$. According to the theory stated in this section we first select $x_{(1)}$. To determine the

remaining seven quantiles we first compute the upper and lower bounds in expressions (3.3), which yield new $\alpha' = 0.40$ and $\beta' = 0.70$. Thus, using Table 1 for these values with $k = 7$ we obtain $k_1 = 2$ and $k_2 = 5$ and optimum spacings as $\lambda^*_{12} = 0.2170$, $\lambda^*_{13} = 0.4000$; $\lambda^*_{21} = 0.7000$, $\lambda^*_{22} = 0.8354$, $\lambda^*_{23} = 0.9226$, $\lambda^*_{24} = 0.9720$, and $\lambda^*_{25} = 0.9943$. Using formula (3.6), we obtain optimum spacings for both (μ, σ) as $p^*_{12} = 0.2295$, $p^*_{13} = 0.4096$, $p^*_{21} = 0.7048$, $p^*_{22} = 0.8381$, $p^*_{23} = 0.9238$, $p^*_{24} = 0.9724$, and $p^*_{25} = 0.9944$. The corresponding ranks of the quantiles are $n_{12} = 15$, $n_{13} = 26$, $n_{21} = 44$, $n_{22} = 52$, $n_{23} = 58$, $n_{24} = 61$, and $n_{25} = 62$. The BLUES are given by

$$\hat{\mu} = x_{(1)} - \hat{\sigma} \ln \frac{125}{123}$$

$$\hat{\sigma} = -.9155x_{(1)} + .2067x_{(15)} + .2774x_{(26)} + .2043x_{(44)} + .1127x_{(52)} + .0681x_{(58)} + .0345x_{(61)} + .0118x_{(62)}.$$

The coefficients b^*_{ij} and b^*_{2j} are taken from Table 1 with $k = 7$.

4. OPTIMUM ALLOCATION OF QUANTILES AND THEIR SPACINGS FOR THE SCALE PARAMETER.

In section 3, we have reduced the two-parameter estimation problem based on k selected quantiles to the problem of estimating the scale parameter based on $k-1$ selected quantiles when the sample is censored in the middle. Therefore, we consider the problem of optimizing the related variance function Q_{k-1} as in (3.2) which is a function of $k-1$ variables. Thus we maximize Q_{k-1} subject to the restrictions

(4.1)
$$\left.\begin{aligned} &\text{(i)} \quad 0 < t_{11} < \ldots < t_{1k_1-1} \leq \ln\left[\frac{n-1/2}{(n+1/2)(1-\alpha)}\right] \\ &\text{and} \\ &\text{(ii)} \quad \ln\left[\frac{n-1/2}{(n+1/2)(1-\beta)}\right] \leq t_{21} \leq \ldots \leq t_{2k_2} < \infty \end{aligned}\right\}.$$

The problem therefore reduces to solving the following system of equations

(4.2)
$$\left.\begin{aligned} \tau_{1i+1} + \tau_{1i} - 2t_{1i} &= 0 \\ \tau_{2j+1} + \tau_{2j} - 2t_{2j} &= 0 \end{aligned}\right\},$$

subject to the restrictions (4.1), where τ_{1i} and τ_{2j} have the same definition as (6.3) of Saleh [5] with additional subscript 1 and 2 in t's.

The theorems in the same paper guarantee that the system of equations (4.2) has a unique solution. Therefore the optimum quantiles for the BLUE of the scale parameter, σ, are uniquely determinable. The nature of the solution depends on the available restrictions and, accordingly, they are as follows:

(i) The solutions coincide with unrestricted optimization problem if the proportion of censoring at the middle is such that

(4.3)
$$t^0_{k_1} \leq \ln\left[\frac{n-1/2}{(n+1/2)(1-\alpha)}\right]$$

and

$$t^0_{1k_1+1} = t^0_{21} \geq \ln\left[\frac{n-1/2}{(n+1/2)(1-\beta)}\right]$$

simultaneously, where $t^0_{1k_1}$ and $t^0_{1k_1+1}$ are the solutions of the equations in (4.2), with no restriction.

(ii) If the solution at (i) is not available, then we proceed as a simultaneous problem of right and left censoring. Accordingly the solution is available following section 3 and 4 of Saleh [5] for (4.2) simultaneously. The associated computation has been performed on a GE 415 Computer with 12-figure accuracy and the iterated solution of the equation has been performed with 5-figure accuracy, for $k=2(1)10$ and $\alpha=0.40(0.10)0.80$ $\beta=0.50(0.10)0.80$, such that $\beta-\alpha=0.10(0.10)0.40$.

The optimum allocation of k, optimum spacings, the coefficient of the BLUE of σ, and the maximum value of Q_{k-1} have been presented at the end of the paper. We mark with an asterisk where the solution is not different from the unrestricted case. The table has been prepared with k instead of $k-1$ to state the result for the scale parameter when the location parameter is known. In the two-parameter case, we use the table for $k-1$ instead of k. The efficiency expression for the BLUE of σ is given by

$$\text{(4.4)} \qquad \text{ARE } (\hat{\sigma}) = \frac{Q^*_{k-1}(\beta-\alpha)}{(\beta-\alpha)(1+\alpha-\beta)+(1-\alpha)(1-\beta)\{\ln(1-\alpha)^{-1}-\ln(1-\beta)^{-1}\}^2}.$$

Now, we shall present an example with finite sample size to illustrate the estimation procedure.

EXAMPLE: Assume $\alpha=0.40$, $\beta=0.60$, $k=7$, $n=60$. From Table 1 we obtain $k_1=1$, $k_2=6$, $\lambda^*_{11}=0.4000$, $\lambda^*_{12}=0.6088$, $\lambda^*_{21}=0.7625$, $\lambda^*_{22}=0.8697$, $\lambda^*_{23}=0.9387$, $\lambda^*_{24}=0.9778$, and $\lambda^*_{25}=0.9955$.
The corresponding order statistics are 25, 37, 46, 53, 57, 59, and 60.
The BLUE of σ is

$$\hat{\sigma}=0.2949x_{(25)}+0.1851x_{(37)}+0.1327x_{(46)}+0.0889x_{(53)}+0.0537x_{(57)}+0.0272x_{(59)}+0.0093x_{(60)}.$$

ARE $(\hat{\sigma})=97.04\%$

5. SOME REMARKS ON THE SIMULTANEOUS ESTIMATION OF μ AND σ BASED ON OPTIMUM QUANTILES

The simultaneous estimation of μ and σ depends heavily on the solution of the scale-parameter problem discussed in section 4 of this paper. The example cited at the end of section 2 illustrates the estimation procedure with associated calculations needed to arrive at the right results. Efficiency expressions are based on the asymptotic approximations of the variances and generalized variance in the finite sample case (Sarhan and Greenberg [7]). Therefore, if the sample size is reasonably large to justify asymptotic normality of the quantiles, the asymptotic efficiencies will also be justified. Finally, the coefficients in the estimation for the scale-parameter case remain the same in the two-parameter case, as well, due to the linear transformations in (3.1) and the nature of the expressions for the coefficients (2.6a to 2.6d). In this regard, the reader is referred to the papers [4, 5] of the primary author, where all details have been given.

ACKNOWLEDGMENT

We are grateful to the referee for his comments and for pointing out Ref. [8], which has helped us to rewrite the paper in the present form.

REFERENCES

[1] Kulldorff, G., "On the Optimum Spacings of the Sample Quantiles from an Exponential Distribution, Final Mimeographed Report," University of Lund, Sweden (1963).

[2] Kulldorff, G., "Estimation of One or Two Parameters of Exponential Distribution on the Basis of Suitably Chosen Order Statistics," Ann. Math. Stat. **34**, 1419–1431 (1963).

[3] Ogawa, J., "Determination of Optimum Spacings for the Estimation of the Scale Parameters of an Exponential Distribution Based on Sample Quantiles," Ann. Inst. Statist. Math. (Tokyo) **12**, 141–155 (1960).

[4] Saleh, A. K. Md. Ehsanes and M. M. Ali, "Asymptotic Optimum Quantiles for the Estimation of the Parameters of the Negative Exponential Distribution," Ann. Math. Statist. **37**, 143–151 (1966).

[5] Saleh, A. K. Md. Ehsanes, "Estimation of the Parameters of the Exponential Distribution Based on Optimum Order Statistics in Censored Samples," Ann. Math. Statist. **37**, 1717–1735 (1966).

[6] Saleh, A. K. Md. Ehsanes, "Determination of Exact Optimum Order Statistics for Estimating the Parameters of the Exponential Distribution in Censored Samples," Technometrics **9**, 279–292 (1967).

[7] Sarhan, A. E., and B. Greenberg (editors) Contribution to Order Statistics (Wiley, New York, 1962), 357–360.

[8] Sarhan, A. E., and B. Greenberg, "Linear Estimates for Doubly Censored Samples from the Exponential Distribution with Observations also Missing from the Middle," Bulletin of the International Statistical Institute, 36th Session, **42**, Book 2 (1967), 1195–1204.

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*

($\beta = 0.50$ $\alpha = 0.40$)

k	2*	3*	4	5*	6*	7*	8*	9	10
k_1	0	0	1	1	1	1	1	2	2
k_2	2	3	3	4	5	6	7	7	8
t_1	1.0176	0.7540	0.5108	0.4993	0.4276	0.3740	0.3324	0.2446	0.2446
λ_1	0.6385	0.5295	0.4000	0.3931	0.3479	0.3120	0.2828	0.2170	0.2170
b_1	0.5232	0.4477	0.3934	0.3463	0.3109	0.2820	0.2579	0.2035	0.2027
t_2	2.6112	1.7716	1.2649	1.0998	0.9269	0.8016	0.7063	0.5108	0.5108
λ_2	0.9266	0.8299	0.7177	0.6671	0.6042	0.5514	0.5066	0.4000	0.4000
b_2	0.1791	0.2266	0.2585	0.2320	0.2228	0.2120	0.2010	0.1926	0.1807
t_3		3.3653	2.2825	1.8539	1.5274	1.3009	1.1339	0.8848	0.8432
λ_3		0.9654	0.8980	0.8434	0.7829	0.7277	0.6782	0.5872	0.5697
b_3		0.0776	0.1308	0.1402	0.1492	0.1519	0.1511	0.1674	0.1535
t_4			3.8761	2.8714	2.2815	1.9014	1.6333	1.3124	1.2172
λ_4			0.9793	0.9434	0.8979	0.8506	0.8047	0.7308	0.7039
b_4			0.0448	0.0709	0.0902	0.1017	0.1083	0.1259	0.1196
t_5				4.4651	3.2990	2.6554	2.2337	1.8118	1.6448
λ_5				0.9885	0.9631	0.9297	0.8929	0.8366	0.8069
b_5				0.0243	0.0456	0.0615	0.0725	0.0902	0.0899
t_6					4.8927	3.6730	2.9878	2.4122	2.1441
λ_6					0.9925	0.9746	0.9496	0.9104	0.8828
b_6					0.0156	0.0311	0.0438	0.0604	0.0644
t_7						5.266	4.0054	3.1663	2.7446
λ_7						0.9948	0.9819	0.9578	0.9357
b_7						0.0106	0.0222	0.0365	0.0432
t_8							5.5990	4.1838	3.4986
λ_8							0.9963	0.9848	0.9698
b_8							0.0076	0.0185	0.0261
t_9								5.7775	4.5162
λ_9								0.9969	0.9891
b_9								0.0063	0.0132
t_{10}									6.1098
λ_{10}									0.9978
b_{10}									0.0045
Q_k	0.8203	0.8910	0.9260	0.9476	0.9606	0.9693	0.9754	0.9794	0.9831

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.40$ $\beta = 0.60$)

k	2*	3	4	5*	6*	7	8	9	10
k_1	0	0	1	1	1	1	2	2	2
k_2	2	3	3	4	5	6	6	7	8
t_1	1.0176	0.9163	0.5108	0.4993	0.4276	0.5108	0.2446	0.2446	0.2446
λ_1	0.6385	0.6000	0.4000	0.3931	0.3479	0.4000	0.2170	0.2170	0.2170
b_1	0.5232	0.4285	0.3934	0.3463	0.3109	0.2949	0.2046	0.2035	0.2028
t_2	2.6112	1.9339	1.2649	1.0998	0.9269	0.9384	0.5108	0.5108	0.5108
λ_2	0.9266	0.8554	0.7177	0.6671	0.6042	0.6088	0.4000	0.4000	0.4000
b_2	0.1791	0.1933	0.2585	0.2320	0.2228	0.1851	0.2079	0.2010	0.2003
t_3		3.5275	2.2825	1.8539	1.5274	1.4378	0.9384	0.9163	0.9163
λ_3		0.9706	0.8980	0.8434	0.7829	0.7625	0.6088	0.6000	0.6000
b_3		0.0662	0.1308	0.1402	0.1492	0.1327	0.1839	0.1695	0.1594
t_4			3.8761	2.8714	2.2815	2.0282	1.4378	1.3439	1.2903
λ_4			0.9793	0.9434	0.8979	0.8697	0.7625	0.7392	0.7248
b_4			0.0448	0.0709	0.0902	0.0889	0.1318	0.1220	0.1112
t_5				4.4651	3.2990	2.7923	2.0382	1.8432	1.7179
λ_5				0.9885	0.9631	0.9387	0.8697	0.8417	0.8206
b_5				0.0243	0.0456	0.0537	0.0883	0.0874	0.0836
t_6					4.8927	3.8099	2.7923	2.4437	2.2172
λ_6					0.9925	0.9778	0.9387	0.9132	0.8911
b_6					0.0156	0.0272	0.0533	0.0585	0.0599
t_7						5.4035	3.8099	3.1977	2.8177
λ_7						0.9955	0.9788	0.9591	0.9403
b_7						0.0093	0.0270	0.0354	0.0401
t_8							5.4035	4.2153	3.5717
λ_8							0.9955	0.9852	0.9719
b_8							0.0092	0.0179	0.0242
t_9								5.8090	4.5893
λ_9								0.9970	0.9898
b_9								0.0061	0.0123
t_{10}									6.1829
λ_{10}									0.9979
b_{10}									0.0042
Q_k	0.8203	0.8878	0.9260	0.9476	0.9606	0.9678	0.9742	0.9794	0.9828

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.40$ $\beta = 0.70$)

k	2	3	4	5	6	7	8	9	10
k_1	0	1	1	1	1	2	2	2	2
k_2	2	2	3	4	5	5	6	7	8
t_1	1.2040	0.5108	0.5108	0.5108	0.5108	0.2446	0.2446	0.2446	0.2446
λ_1	0.7000	0.4000	0.4000	0.4000	0.4000	0.2170	0.2170	0.2170	0.2170
b_1	0.4832	0.4750	0.3934	0.3700	0.3658	0.2067	0.2054	0.2045	0.2040
t_2	2.7976	1.5284	1.2649	1.2040	1.2040	0.5108	0.5108	0.5108	0.5108
λ_2	0.9390	0.7831	0.7177	0.7000	0.7000	0.4000	0.4000	0.4000	0.4000
b_2	0.1495	0.2914	0.2585	0.2269	0.2269	0.2774	0.2757	0.2746	0.2738
t_3		3.1220	2.2825	1.9580	1.8044	1.2040	1.2040	1.2040	1.2040
λ_3		0.9559	0.8980	0.8589	0.8354	0.7000	0.7000	0.7000	0.7000
b_3		0.0997	0.1380	0.1264	0.1134	0.2043	0.1902	0.1801	0.1725
t_4			3.8761	2.9756	2.5585	1.8044	1.7033	1.6316	1.5780
λ_4			0.9793	0.9490	0.9226	0.8354	0.8179	0.8044	0.7936
b_4			0.0448	0.0640	0.0685	0.1127	0.1015	0.0920	0.0839
t_5				4.5692	3.5761	2.5585	2.3038	2.1309	2.0056
λ_5				0.9896	0.9720	0.9226	0.9001	0.8813	0.8654
b_5				0.0219	0.0347	0.0681	0.0680	0.0659	0.0631
t_6					5.1697	3.5761	3.0578	2.7314	2.5019
λ_6					0.9943	0.9720	0.9530	0.9349	0.9183
b_6					0.0119	0.0345	0.0411	0.0441	0.0452
t_7						5.1697	4.0754	3.4854	3.1054
λ_7						0.9943	0.9830	0.9694	0.9552
b_7						0.0118	0.0208	0.0267	0.0303
t_8							5.6690	4.5030	3.8594
λ_8							0.9965	0.9889	0.9789
b_8							0.0071	0.0135	0.0183
t_9								6.0966	4.8770
λ_9								0.9977	0.9924
b_9								0.0046	0.0093
t_{10}									6.4706
λ_{10}									0.9985
b_{10}									0.0032
Q_k	0.8155	0.8836	0.9260	0.9470	0.9578	0.9642	0.9704	0.9743	0.9769

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\beta = 0.80$ $\alpha = 0.40$)

k	2	3	4	5	6	7	8	9	10
k_1	1	1	1	1	1	2	2	2	2
k_2	1	2	3	4	5	5	6	7	8
t_1	0.5108	0.5108	0.5108	0.5108	0.5108	0.2446	0.2446	0.2446	0.2446
λ_1	0.4000	0.4000	0.4000	0.4000	0.4000	0.2170	0.2170	0.2170	0.2170
b_1	0.6698	0.4945	0.4759	0.4687	0.4651	0.2108	0.2099	0.2093	0.2089
t_2	2.1045	1.6094	1.6094	1.6094	1.6094	0.5108	0.5108	0.5108	0.5180
λ_2	0.8781	0.8000	0.8000	0.8000	0.8000	0.4000	0.4000	0.4000	0.4000
b_2	0.3126	0.2812	0.2336	0.2099	0.1956	0.3743	0.3727	0.3716	0.3710
t_3		3.2031	2.6270	2.3635	2.2029	1.6094	1.6094	1.6094	1.6094
λ_3		0.9594	0.9277	0.9059	0.8903	0.8000	0.8000	0.8000	0.8000
b_3		0.0920	0.0935	0.0856	0.0771	0.1943	0.1847	0.1778	0.1726
t_4			4.2207	3.3811	2.9639	2.2099	2.1088	2.0370	1.9834
λ_4			0.9853	0.9660	0.9484	0.8903	0.8786	0.8696	0.8624
b_4			0.0320	0.0433	0.0466	0.0766	0.0691	0.0627	0.0573
t_5				4.9747	3.9815	2.9639	2.7092	2.5364	2.4110
λ_5				0.9931	0.9813	0.9484	0.9334	0.9208	0.9103
b_5				0.0148	0.0236	0.0463	0.0463	0.0450	0.0431
t_6					5.5752	3.9815	3.4633	3.1368	2.9104
λ_6					0.9962	0.9813	0.9687	0.9566	0.9455
b_6					0.0081	0.0234	0.0280	0.0301	0.0309
t_7						5.5752	4.4809	3.8909	3.5108
λ_7						0.9962	0.9887	0.9796	0.9701
b_7						0.0080	0.0142	0.0182	0.0207
t_8							6.0745	4.9085	4.2649
λ_8							0.9977	0.9926	0.9859
b_8							0.0048	0.0092	0.0125
t_9								6.5021	5.2825
λ_9								0.9985	0.9949
b_9								0.0032	0.0063
t_{10}									6.8761
λ_{10}									0.9990
b_{10}									0.0022
Q_k	0.7800	0.8830	0.9176	0.9317	0.9389	0.9453	0.9494	0.9520	0.9538

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.50$ $\beta = 0.60$)

k	2*	3	4*	5*	6*	7	8	9*	10
k_1	0	1	1	1	1	2	2	2	2
k_2	2	2	3	4	5	5	6	7	8
t_1	1.0176	0.6931	0.6005	0.4993	0.4276	0.3266	0.3266	0.2991	0.3266
λ_1	0.6385	0.5000	0.4514	0.3931	0.3479	0.2786	0.2786	0.2585	0.2786
b_1	0.5232	0.4519	0.3907	0.3463	0.3109	0.2563	0.2546	0.2376	0.2527
t_2	2.6112	1.7107	1.3545	1.0998	0.9269	0.6931	0.6931	0.6315	0.6931
λ_2	0.9266	0.8193	0.7419	0.6671	0.6042	0.5000	0.5000	0.4682	0.5000
b_2	0.1791	0.2409	0.2361	0.2320	0.2228	0.2185	0.2015	0.1904	0.1788
t_3		3.3044	2.3721	1.8539	1.5274	1.1925	1.1207	1.0055	1.0255
λ_3		0.9633	0.9067	0.8434	0.7829	0.6965	0.6740	0.6341	0.6414
b_3		0.0825	0.1195	0.1402	0.1492	0.1694	0.1531	0.1483	0.1280
t_4			3.9657	2.8714	2.2815	1.7929	1.6201	1.4331	1.3995
λ_4			0.9810	0.9434	0.8979	0.8335	0.8021	0.7614	0.7533
b_4			0.0409	0.0709	0.0902	0.1134	0.1097	0.1115	0.0997
t_5				4.4651	3.2990	2.5470	2.2206	1.9324	1.8271
λ_5				0.9885	0.9631	0.9217	0.8915	0.8552	0.8391
b_5				0.0243	0.0456	0.0685	0.0735	0.0799	0.0750
t_6					4.8927	3.5646	2.9746	2.5329	2.3264
λ_6					0.9925	0.9717	0.9489	0.9206	0.9024
b_6					0.0156	0.0347	0.0444	0.0535	0.0537
t_7						5.1582	3.9922	3.2869	2.9269
λ_7						0.9942	0.9815	0.9626	0.9464
b_7						0.0119	0.0225	0.0323	0.0360
t_8							5.5858	4.3045	3.6809
λ_8							0.9962	0.9865	0.9748
b_8							0.0077	0.0164	0.0217
t_9								5.8981	4.6985
λ_9								0.9973	0.9909
b_9								0.0056	0.0110
t_{10}									6.2922
λ_{10}									0.9981
b_{10}									0.0038
Q_k	0.8203	0.8906	0.9269	0.9476	0.9606	0.9689	0.9754	0.9798	0.9828

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.50$ $\beta = 0.70$)

n	2	3	4*	5	6	7	8	9	10
k_1	0	1	1	1	2	2	2	2	3
k_2	2	2	3	4	4	5	6	7	7
t_1	1.2040	0.6931	0.6005	0.6931	0.3266	0.3266	0.3266	0.3266	0.2138
λ_1	0.7000	0.5000	0.4514	0.5000	0.2786	0.2786	0.2786	0.2786	0.1925
b_1	0.4832	0.4549	0.3907	0.3478	0.2591	0.2563	0.2547	0.2537	0.1821
t_2	2.7976	1.7107	1.3545	1.2935	0.6931	0.6931	0.6931	0.6931	0.4439
λ_2	0.9390	0.8193	0.7419	0.7257	0.5000	0.5000	0.5000	0.5000	0.3585
b_2	0.1495	0.2409	0.2361	0.1918	0.2425	0.2210	0.2196	0.2187	0.1561
t_3		3.3044	2.3721	2.0477	1.2936	1.2040	1.2040	1.2040	0.6931
λ_3		0.9633	0.9067	0.8710	0.7257	0.7000	0.7000	0.7000	0.5000
b_3		0.0825	0.1195	0.1159	0.1889	0.1695	0.1557	0.1458	0.1852
t_4			3.9657	3.0652	2.0477	1.8044	1.7033	1.6316	1.2040
λ_4			0.9810	0.9534	0.8710	0.8354	0.8179	0.8044	0.7000
b_4			0.0409	0.0587	0.1141	0.1121	0.1010	0.0915	0.1454
t_5				4.6589	3.0652	2.5585	2.3038	2.1309	1.6316
λ_5				0.9905	0.9534	0.9226	0.9001	0.8813	0.8044
b_5				0.0201	0.0578	0.0678	0.0676	0.0656	0.0913
t_6					4.6589	3.5761	3.0578	2.7314	2.1309
λ_6					0.9905	0.9720	0.9530	0.9349	0.8813
b_6					0.0193	0.0343	0.0409	0.0439	0.0654
t_7						5.1697	4.0754	3.4854	2.7314
λ_7						0.9943	0.9830	0.9694	0.9349
b_7						0.0117	0.0207	0.0265	0.0438
t_8							5.6690	4.5030	3.4854
λ_8							0.9965	0.9889	0.9694
b_8							0.0071	0.0134	0.0265
t_9								6.0966	4.5030
λ_9								0.9977	0.9889
b_9								0.0046	0.0134
t_{10}									6.0966
λ_{10}									0.9977
b_{10}									0.0046
Q_k	0.8155	0.8906	0.9269	0.9439	0.9585	0.9689	0.9751	0.9789	0.9817

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.50$ $\beta = 0.80$)

k	2	3	4	5	6	7	8	9	10
k_1	1	1	1	2	2	2	2	3	3
k_2	1	2	3	3	4	5	6	6	7
t_1	0.6931	0.6931	0.6931	0.3266	0.3266	0.3266	0.3266	0.2138	0.2138
λ_1	0.5000	0.5000	0.5000	0.2786	0.2786	0.2786	0.2786	0.1925	0.1925
b_1	0.6092	0.4549	0.4194	0.2645	0.2606	0.2586	0.2575	0.1848	0.1843
t_2	2.2868	1.7107	1.6094	0.6931	0.6931	0.6931	0.6931	0.4439	0.4439
λ_2	0.8984	0.8193	0.8000	0.5000	0.5000	0.5000	0.5000	0.3585	0.3585
b_2	0.2526	0.2409	0.2058	0.3108	0.3061	0.3039	0.3025	0.1585	0.1580
t_3		3.3044	2.6270	1.6094	1.6094	1.6094	1.6094	0.6931	0.6931
λ_3		0.9633	0.9277	0.8000	0.8000	0.8000	0.8000	0.5000	0.5000
b_3		0.0825	0.9029	0.2026	0.1799	0.1661	0.1568	0.2683	0.2675
t_4			4.2207	2.6270	2.3634	2.2099	2.1088	1.6094	1.6094
λ_4			0.9853	0.9277	0.9059	0.8903	0.8786	0.8006	0.8000
b_4			0.0318	0.0914	0.0837	0.0754	0.0681	0.1564	0.1497
t_5				4.2207	3.3811	2.9639	2.7092	2.1088	2.0370
λ_5				0.9853	0.9660	0.9484	0.9334	0.8786	0.8696
b_5				0.0313	0.0424	0.0456	0.0456	0.0679	0.0616
t_6					4.9747	3.9815	3.4633	2.7092	2 5364
λ_6					0.9931	0.9813	0.9687	0.9334	0.9208
b_6					0.0145	0.0231	0.0275	0.0455	0.0441
t_7						5.5752	4.4809	3.4633	3.1368
λ_7						0.9962	0.9887	0.9687	0.9566
b_7						0.0079	0.0139	0.0275	0.0296
t_8							6.07 5	4.4809	3.8909
λ_8							0.9977	0.9889	0.9796
b_8							0.0048	0.0139	0.0179
t_9								6.0745	4.9085
λ_9								0.9977	0.9926
b_9								0.0048	0.0090
t_{10}									6.5021
λ_{10}									0.9985
b_{10}									0.0031
Q_k	0.8043	0.8906	0.9244	0.9390	0.9531	0.9603	0.9645	0.9672	0.9698

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.60$ $\beta = 0.70$)

k	2	3*	4*	5	6	7*	8	9	10*
k_1	1	1	1	2	2	2	3	3	3
k_2	1	2	3	3	4	5	5	6	7
t_1	0.9163	0.7540	0.6005	0.4232	0.4232	0.3740	0.2755	0.2755	0.2719
λ_1	0.6000	0.5295	0.4514	0.3450	0.3450	0.3120	0.2408	0.2408	0.2381
b_1	0.5475	0.4477	0.3907	0.3135	0.3088	0.2820	0.2244	0.2232	0.2203
t_2	2.5099	1.7716	1.3545	0.9163	0.9163	0.8016	0.5788	0.5788	0.5711
λ_2	0.9187	0.8299	0.7419	0.6000	0.6000	0.5514	0.4394	0.4394	0.4351
b_2	0.1985	0.2266	0.2361	0.2523	0.2237	0.2120	0.1833	0.1823	0.1804
t_3		3.3653	2.3721	1.6703	1.5167	1.3099	0.9163	0.9163	0.9034
λ_3		0.9654	0.9067	0.8118	0.7806	0.7277	0.6000	0.6000	0.5948
b_3		0.0776	0.1195	0.1686	0.1508	0.1519	0.1671	0.1539	0.1445
t_4			3.9657	2.6879	2.2708	1.9014	1.4156	1.3439	1.2774
λ_4			0.9810	0.9320	0.8968	0.8506	0.7572	0.7392	0.7212
b_4			0.0409	0.0854	0.0911	0.1017	0.1347	0.1219	0.1126
t_5				4.2816	3.2834	2.6554	2.0161	1.8432	1.7050
λ_5				0.9862	0.9627	0.9297	0.8668	0.8417	0.8182
b_5				0.0292	0.0461	0.0615	0.0902	0.0874	0.0847
t_6					4.8820	3.6730	2.7701	2.4427	2.2043
λ_6					0.9924	0.9746	0.9373	0.9132	0.8897
b_6					0.0158	0.0311	0.0545	0.0585	0.0607
t_7						5.2666	3.7877	3.1977	2.8048
λ_7						0.9948	0.9774	0.9591	0.9395
b_7						0.0106	0.0276	0.0354	0.0406
t_8							5.3814	4.2153	3.5588
λ_8							0.9954	0.9852	0.9715
b_8							0.0094	0.0179	0.0246
t_9								5.8090	4.5764
λ_9								0.9970	0.9897
b_9								0.0061	0.0124
t_{10}									6.1701
λ_{10}									0.9979
b_{10}									0.0043
Q_k	0.8188	0.8910	0.9269	0.9462	0.9606	0.9693	0.9745	0.9797	0.9832

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.60$ $\beta = 0.80$)

k	2	3*	4	5	6	7	8	9	10
k_1	1	1	2	2	2	2	3	3	3
k_2	1	2	2	3	4	5	5	6	7
t_1	0.9163	0.7540	0.4232	0.4232	0.4232	0.4232	0.2755	0.2755	0.2755
λ_1	0.6000	0.5295	0.3450	0.3450	0.3450	0.3450	0.2408	0.2408	0.2408
b_1	0.5475	0.4477	0.3231	0.3135	0.3089	0.3066	0.2248	0.2238	0.2232
t_2	2.5099	1.7716	0.9163	0.9163	0.9163	0.9163	0.5788	0.5788	0.5788
λ_2	0.9187	0.8299	0.6000	0.6000	0.6000	0.6000	0.4394	0.4394	0.4394
b_2	0.1985	0.2266	0.3010	0.2523	0.2389	0.2372	0.1835	0.1828	0.1823
t_3		3.3653	1.9239	1.6703	1.6094	1.6094	0.9163	0.9163	0.9163
λ_3		0.9654	0.8154	0.8118	0.8000	0.8000	0.6000	0.6000	0.6000
b_3		0.0774	0.1870	0.1686	0.1492	0.1358	0.1994	0.1986	0.1980
t_4			3.5275	2.6879	2.3635	2.2099	1.6094	1.6094	1.6094
λ_4			0.9706	0.9320	0.9059	0.8903	0.8000	0.8000	0.8000
b_4			0.0640	0.0854	0.0831	0.0749	0.1340	0.1259	0.1194
t_5				4.2816	3.3811	2.9639	2.2099	2.1088	2.0370
λ_5				0.9862	0.9660	0.9484	0.8903	0.8786	0.8696
b_5				0.0292	0.0421	0.0452	0.0744	0.0672	0.0610
t_6					4.9747	3.9815	2.9639	2.7092	2.5364
λ_6					0.9931	0.9813	0.9484	0.9334	0.9208
b_6					0.0144	0.0229	0.0450	0.0450	0.0437
t_7						5.5752	3.9815	3.4633	3.1368
λ_7						0.9962	0.9813	0.9687	0.9566
b_7						0.0078	0.0228	0.0272	0.0293
t_8							5.5752	4.4809	3.8909
λ_8							0.9962	0.9887	0.9796
b_8							0.0078	0.0138	0.0177
t_9								6.0745	4.9085
λ_9								0.9977	0.9926
b_9								0.0047	0.0089
t_{10}									6.5021
λ_{10}									0.9958
b_{10}									0.0031
Q_k	0.8188	0.8910	0.9179	0.9462	0.9502	0.9674	0.9730	0.9771	0.9797

TABLE 1. *Optimum Spacings, the Corresponding Coefficients, and Relative Efficiency of the Scale Parameter*—Continued

($\alpha = 0.70$ $\beta = 0.80$)

k	2*	3*	4	5*	6	7	8*	9	10
k_1	1	1	2	2	3	3	3	4	4
k_2	1	2	2	3	3	4	5	5	6
t_1	1.0176	0.7540	0.5414	0.4993	0.3501	0.3501	0.3324	0.2588	0.2588
λ_1	0.6385	0.5295	0.4181	0.3931	0.2954	0.2954	0.2828	0.2280	0.2280
b_1	0.5232	0.4477	0.3708	0.3463	0.2724	0.2694	0.2579	0.2128	0.2119
t_2	2.6112	1.7716	1.2040	1.0998	0.7467	0.7467	0.7063	0.5420	0.5420
λ_2	0.9266	0.8299	0.7000	0.6671	0.5261	0.5261	0.5066	0.4184	0.4184
b_2	0.1791	0.2266	0.2565	0.2320	0.2090	0.2067	0.2010	0.1761	0.1754
t_3		3.3653	2.2216	1.8539	1.2040	1.2040	1.1339	0.8548	0.8548
λ_3		0.9654	0.8916	0.8434	0.7000	0.7000	0.6782	0.5746	0.5746
b_3		0.0776	0.1390	0.1402	0.1804	0.1804	0.1511	0.1429	0.1423
t_4			3.8152	2.8714	1.9580	1.8044	1.6333	1.2040	1.2040
λ_4			0.9780	0.9434	0.8589	0.8354	0.8047	0.7000	0.7000
b_4			0.0476	0.0709	0.1249	0.1121	0.1083	0.1268	0.1170
t_5				4.4651	2.9756	2.5585	2.2337	1.7033	1.6316
λ_5				0.9885	0.9490	0.9226	0.8929	0.8179	0.8044
b_5				0.0243	0.0632	0.0677	0.0725	0.1005	0.9011
t_6					4.5692	3.5761	2.9878	2.3038	2.1309
λ_6					0.9896	0.9720	0.9496	0.9001	0.8813
b_6					0.0216	0.0343	0.0438	0.0673	0.0653
t_7						5.1697	4.0054	3.0578	2.7314
λ_7						0.9943	0.9818	0.9530	0.9349
b_7						0.0117	0.0222	0.0407	0.0437
t_8							5.5990	4.0754	3.4854
λ_8							0.9963	0.9830	0.9694
b_8							0.0076	0.0206	0.0264
t_9								5.6690	4.5030
λ_9								0.9965	0.9889
b_9								0.0070	0.0134
t_{10}									6.0966
λ_{10}									0.9977
b_{10}									0.0046
Q_k	0.8203	0.8910	0.9259	0.9476	0.9583	0.9691	0.9754	0.9792	0.9831

DECISION RULES FOR EQUAL SHORTAGE POLICIES

G. Gerson

Cambridge Computer Corp.
New York, N.Y.

and

R. G. Brown

IBM Corporation
White Plains, N.Y.

I. INTRODUCTION

Much of the applied work in inventory management has been based on "equal service" policies—i.e., each item in an inventory should be managed in such a way that over a year the same percentage dollar demand for the item can be met.

This paper presents a set of practical decision rules for "equal shortage" policies—i.e., each item in an inventory should have the same number of shortage occurrences in the course of a year. It also answers the question of allocating inventories under budgetary constraints.

There is a substantial difference between the two policies of "equal service" and "equal shortage." If one aims for a desired level of service, in terms of dollar demand filled from the shelf, presumably the number of shortages are not of paramount importance—and conversely. A total inventory budget is allocated among the items in the inventory in quite different ways under the two policies.

There can also be a strategic problem in allocating an inventory under a budgetary constraint. With a fixed amount of cash available for inventory (or an equivalent measure of value, such as shelf space available), what safety factors should be used in computing buffer stocks, and what ordering quantities should be used to:

a. Yield minimum dollar shortages for the inventory (in terms of lost demand), or

b. Yield minimum number of shortage occurrences.

In this paper the decision rules developed will meet budgetary constraints and allocate the inventory so as to satisfy either *a* or *b*. It is also shown in the development that the way to meet a budget and satisfy *a* is to invoke the "equal shortages" policy, contrary to policies implemented in IMPACT, for example, which concentrate on "equal service" rules.

In Section II we develop the decision rules required where every stock item is ordered with the the same frequency, as is often the case for retailers and wholesalers. In Section III we develop the decision rules in the case where each item may have its own ordering frequency. In Section IV ordering and holding costs are considered in order to minimize total expense under a given capital budget.

II. FIXED ORDERING FREQUENCIES

In this section we deal with the case where each item is ordered with a known frequency, and shortages are backordered.

The notation to be used is found in Brown [1] and is as follows:

Let $x(t)$ represent the number of units of a given item that is demanded at time t. Assume that $x(t)$ has mean $\bar{x}$ and standard deviation σ. We also define the deviation at time t, e_t, by

$$e_t = x(t) - \bar{x}.$$

Then e_t has mean 0 and standard deviation σ. Let $p(t)$ be the p.d.f. of e_t. Define

$$F(k) = \int_k^{\infty} p(t)dt.$$

$F(k)$ is the complement of the usual cumulative distribution function, and represents the probability that demand will exceed $\bar{x} + k\sigma$. Define

$$E(k) = \int_k^{\infty} (t-k)p(t)dt.$$

This function is called the "Partial Expectation." The quantity $\sigma E(k)$ represents the expected quantity short per order cycle. Let S represent the annual sales of an item and Q the order quantity. Let v represent the unit value of an item—although v may also be considered in terms of square feet taken up by the item in a shelf allocation procedure. Then S/Q is the number of order cycles in a year. Since $F(k)$ represents the probability that demand will exceed $\bar{x} + k\sigma$, then $F(k)S/Q$ will represent the expected number of shortage occurrences in a year—i.e., the expected number of times in which an out-of-stock situation will occur. With v defined as the unit value of the item, then $\sigma v E(k)S/Q$ will represent the expected dollar value of the shortages.

Throughout the development we consider an inventory investment of the form

$$I = \sum_{j=1}^{n} (k_j \sigma_j v_j + Q_j v_j/2).$$

Consider a fixed budget for I. Then $k_j \sigma_j v_j$ represents safety stock for the jth item, and $Q_j v_j/2$ is the value of cycle stock for the jth item.

THEOREM 1: Given an inventory of n items, dollar shortages will be minimized when all items have the same number of shortage occurrences per year. In particular, if all items are reordered with the same frequency, then all safety factors, k_j, should be equal.

PROOF: Consider the individual values of cycle stock to be fixed. Fix the total investment in safety stocks as I_s. Thus,

$$I_s = \sum_{j=1}^{n} k_j \sigma_j v_j.$$

Total annual shortages are

$$P = \sum_{j=1}^{n} \sigma_j v_j E(k_j) S_j / Q_j.$$

Form

$$H = P - \lambda \left(I_s - \sum_{j=1}^{n} k_j \sigma_j v_j \right),$$

where λ is a Lagrangian multiplier.

To minimize P subject to the constraint on investment, set $\frac{\partial H}{\partial k_j}=0$, and solve to obtain

$$F(k_j)S_j/Q_j=\lambda. \qquad j=1, \ldots, n.$$

Hence, if safety factors, k_j, are chosen so that each item has the same number, λ, of shortage occurrences per year, the dollar value of the backorders is minimized. In particular, if all items are reordered with the same frequency, i.e.,

$$S_j/Q_j=c.$$

then $F(k_j)=c\lambda$, and all safety factors must be the same.

$$k_j=F^{-1}(c\lambda)^*.$$

The same technique can be applied to find the values of k for which the number of shortages will be minimized (if number of shortages is an appropriate definition of service). The resulting equation is

$$\sigma_j v_j S_j p(k_j)/Q_j=\lambda.$$

In this case a restriction must be made on the form of $p(k)$ in order to assure a unique solution—i.e., $p'(k)<0$.

In this theorem, each value of $0<\lambda<S_j/Q_j$ generates a total value of inventory. By varying λ an exchange curve can be generated that yields shortages as a function of inventory investment.

III. RELAXATION OF THE FREQUENCY CONSTRAINT

In this section we consider the case where order quantities Q_j are to be determined jointly with the safety factors so as to minimize the total value of shortages.

$$T(Q_1, \ldots, Q_n, k_1, \ldots, k_n)=\sum_{j=1}^{n}\sigma_j v_j E(k_j)S_j/Q_j$$

subject to a total inventory budget I.

In order to apply Lagrangian Multiplier techniques, we must be sure that the Hessian of T is positive definite. Define

$$t(Q, k)=E(k)/Q.$$

Evaluating

$$\begin{vmatrix} \frac{\partial^2 t}{\partial k^2} & \frac{\partial^2 t}{\partial x \partial Q} \\ \frac{\partial^2 t}{\partial Q \partial k} & \frac{\partial^2 t}{\partial Q^2} \end{vmatrix}$$

we obtain

$$E(k)[2p(k)-F^2(k)/E(k)]/Q^4.$$

*Although this case may seem artificial, it is common practice in industry to order items based on a fixed number of months' supply.

However, applying L'Hospital's Rule

$$\lim_{k\to\infty} F^2(k)/E(k) = 2p(k).$$

Further, taking the derivative,

$$[2E(k)p(k) - F^2(k)]' = 2E(k)p'(k) < 0 \text{ if } p'(k) < 0.$$

Therefore, the Hessian will be positive definite if $p'(k) < 0$. We will assume from now on that this is the case and that Lagrangian Multiplier techniques can be applied as required.

The graph in Appendix 2 shows how the function $F^2(k)/E(k)$ approaches $2p(k)$ as k increases. In this case, $p(k)$ is the normal density function.

THEOREM 2: Given a total inventory constraint

$$I = \sum_{j=1}^{n} (k_j\sigma_j v_j + Q_j v_j/2),$$

then the value of shortages

$$\sum_{j=1}^{n} S_j\sigma_j v_j E(k_j)/Q_j$$

will be minimized, providing that $p'(k) < 0$ and $0 < \lambda < S_j/Q_j$, if the safety factors satisfy

$$F^2(k_j) = 2\sigma_j\lambda E(k_j)/S_j$$

and the order quantities satisfy

$$Q_j = 2\sigma_j E(k_j)/F(k_j).$$

PROOF: The sum of cycle and safety stocks is

$$I = \sum_{j=1}^{n} (k_j\sigma_j v_j + Q_j v_j/2),$$

and the value of shortages is

$$P = \sum_{j=1}^{n} S_j\sigma_j v_j E(k_j)/Q_j.$$

Form

$$H = P - \lambda\left[I - \sum_{j=1}^{n} (k_j\sigma_j v_j + Q_j v_j/2)\right].$$

Take $\frac{\partial H}{\partial k_j}$ and $\frac{\partial H}{\partial Q_j}$, equate them to 0, and solve to get

(1) $$F^2(k_j) = 2\sigma_j\lambda E(k_j)/S_j$$

and

(2) $$Q_j = 2\sigma_j E(k_j)/F(k_j).$$

Equation (1) can be solved for k_j by a Newton iteration for k_j or table look-up. The specific formula required for the Newton iteration is exhibited in Appendix 1. Once k_j has been determined, then Q_j

can be obtained from Eq. (2). Therefore, given a λ which satisfies the hypotheses, we can determine Q_j and k_j. Since $E(k_j)$ is the expected quantity short per order cycle, the average service P_j is defined by

$$\sigma_j E(k_j) = Q_j(1-P_j). \tag{3}$$

Substitute (3) in (2) to obtain

$$F(k_j) = 2(1-P_j).$$

Hence $P_j > 0.5$, and at least half the value of demand will be satisfied on the average. For the normal distribution this means that the safety factors are nonnegative.

If the number of shortages rather than the value of shortages is the criterion for service, then the development is:

$$I = \sum_{j=1}^{n} (k_j \sigma_j v_j + Q_j v_j/2),$$

and the number of shortages is

$$P = \sum_{j=1}^{n} S_j F(k_j)/Q_j.$$

Set

$$H = P - \lambda \left[I - \sum_{j=1}^{n} (k_j \sigma_j v_j + Q_j v_j/2) \right]$$

Again, solve $\frac{\partial H}{\partial k_j} = 0$ and $\frac{\partial H}{\partial Q_j} = 0$ to get

$$p^2(k_j) = 2\lambda \sigma_j v_j F(k_j)/S_j,$$

and

$$Q_j = 2\sigma_j F(k_j)/p(k_j).$$

If these equations are to supply a feasible solution, then the Hessian of $F(k)/Q$ must be positive definite. The Hessian is

$$-F(k)[2p'(k)F(k) + p^2(k)]/Q^4.$$

The expression in the brackets must be negative. The limit of the expression is 0, but

$$[2p'(k)F(k) + p^2(k)]'$$

will be positive only if the second derivative, $p''(k) > 0$.

IV. ORDERING AND SHORTAGE COSTS

Consider a cost c_j of processing a replenishment order, and an expense u_j for processing each piece backordered. Then the total annual expense is

$$X=\sum_{j=1}^{n} c_jS_j/Q+\sum_{j=1}^{n} u_jS_j\sigma_jE(k_j)/Q_j$$

with total inventory

$$I=\sum_{j=1}^{n} (k_j\sigma_jv_j+Q_jv_j/2).$$

Form

$$H=X-\lambda\left[I-\sum_{j=1}^{n} (k_j\sigma_jv_j+Q_jv_j/2)\right].$$

Then

$$\frac{\partial H}{\partial k_j}=-u_jS_j\sigma_jF(k_j)/Q_j+\lambda\sigma_jv_j=0,$$

and

$$\frac{\partial H}{\partial Q_j}=-(u_jS_j\sigma_jE(k_j)+c_jS_j)/Q^2+\lambda v_j/2=0.$$

The second equation reduces to

$$Q_j=\sqrt{2(u_j\sigma_jE(k_j)+c_jS_j)/\lambda v_j}, \tag{4}$$

which becomes the conventional *EOQ* for $\sigma_j=0$. Note that λ is the policy variable that governs the exchange between capital invested and ordering expense, sometimes called the "carrying charge."

The first equation becomes

$$u_jS_jF(k_j)/v_jQ_j=\lambda, \tag{5}$$

which modifies earlier results only in terms of the ratio of the cost per unit backordered to the cost per unit kept in inventory.

It would also be possible to consider a cost U per backorder processed (i.e., it costs something to process the backorder, but the cost is not dependent on the quantity backordered). Then the results will come out like the minimum shortage case considered earlier.

Numerical Examples:

1. Consider the case where order quantities are fixed and all items are reordered with the same frequency.

$$S_1=100 \qquad S_2=200$$
$$v_1=1 \qquad v_2=2$$
$$Q_1=10 \qquad Q_2=20$$
$$\sigma_1=10 \qquad \sigma_2=5$$

We consider for this example that I is made up of safety stocks alone, since the order quantities are fixed.

Set $I=20$. Then we have $k=1$ for both items. The shortages turn out to be 16.66, and the total service is 0.9667. If we use equal service rules, then k_j is computed from

$$\sigma_jE(k_j)=(1-P_j).$$

Then $k_1 = 1.444$ and $k_2 = 0.74$. The total inventory investment necessary to supply an item service of 0.9667 turns out to be 21.84.

2. Consider the case where order quantities and safety factors are determined jointly.

$$S_1 = 100 \qquad S_2 = 200$$
$$v_1 = 1 \qquad v_2 = 1$$
$$\sigma_1 = 6.1 \qquad \sigma_2 = 3.85$$
$$\lambda = 0.5$$

Then $k_1 = 2.0$ and $k_2 = 2.5$. We also obtain $Q_1 = 4.55$ and $Q_2 = 2.48$. The total shortages turn out to be $0.621 + 1.138 = 1.759$. Service for the two items is 0.99414, and the total inventory investment is 25.34. If we consider the equal service strategy with order quantities as above, then $k_1 = 2.239$ and $k_2 = 2.290$. The inventory investment is therefore 25.98.

If we leave $k_1 = 2.0$ and $k_2 = 2.5$, and determine Q's to get an equal service strategy, then $Q_1 = 8.839$ and $Q_2 = 1.214$, so that the total inventory investment turns out to be 26.85.

REFERENCES

[1] Brown, R. G., *Decision Rules for Inventory Management* (Holt, Rhinehart, and Winston, New York, N.Y., 1967).

[2] Hadley, G. and Whitin, T. M., *Analysis of Inventory Systems* (Prentice Hall, Englewood Cliffs, N.J., 1963).

APPENDIX 1

The general Newton iteration method is, with k_0 chosen in advance

$$k_{i+1} = k_i - f(k_i)/f'(k_i).$$

In this case—Eq. (1)—we set

$$c_j = 2\sigma_j 2/S_j.$$

Then

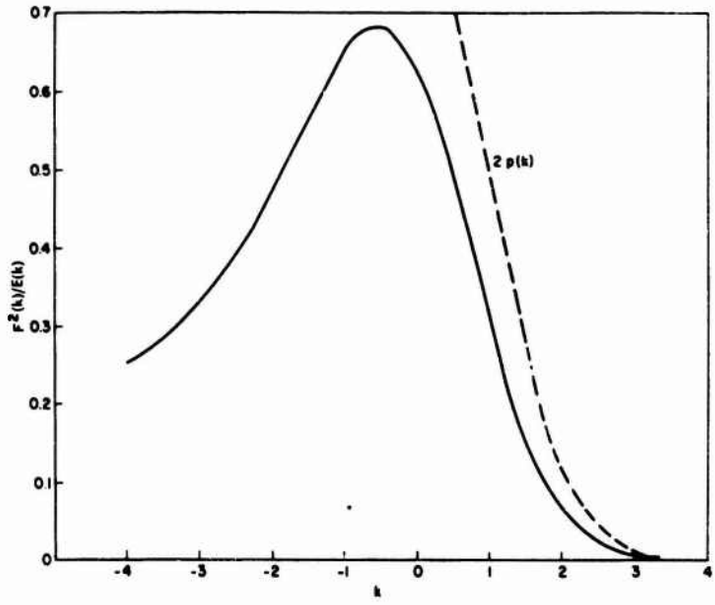
$$f(k_i) = F^2(k_i)/E(k_i) - c_j$$

and

$$f'(k_i) = F(k_i)[F^2(k_i) - 2E(k_i)p(k_i)]/E^2(k_i).$$

The nonvanishing of $f'(k_i)$ is assured by the fact that the Hessian is positive definite. This procedure has been programmed and convergence is rapid.

APPENDIX 2



Approximation of $F^2(k)/E(k)$ by $2p(k)$ where $p(k)$ is the normal density

SYSTEMS ANALYSIS AND PLANNING-PROGRAMMING-BUDGETING SYSTEMS (PPBS) FOR DEFENSE DECISION MAKING

Richard L. Nolan*

Harvard University

ABSTRACT

Systems analysis office titles have permeated both government and business organization charts in recent years. Systems analysis as a discipline, however, even though increasingly accepted, has eluded precise definition. For the most part, it has been loosely described as "quantitative common sense" and "the general application of the scientific method." Emphasis is placed upon the application of eclectic disciplines to a wide variety of problems. Concepts and techniques have been drawn heavily from economics, mathematics, and political science.

In the Department of Defense, systems analysis has been used extensively in the evaluation of weapon systems during the last 9 years. During the 1960's, it provided the underlying concepts for the control system PPBS (Planning-Programming-Budgeting System). This article traces the origins of systems analysis within the Department of Defense and describes and analyzes the application of the technique. Although there always exists disagreement, it is generally accepted that the origin of systems analysis coincided with the inception of R. S. McNamara's administration of the Department of Defense. McNamara organized the Systems Analysis office under Mr. Charles Hitch, who had previously developed many basic systems analysis concepts at project RAND. From Hitch's basic concepts, the approach became increasingly sophisticated in evaluating complex weapons systems. Coincidently, the organizational procedures for implementing systems analysis also evolved. Under the current Department of Defense administration, the new organizational procedures emerging are contrasted with the old.

The allocation of resources for national security must always compete for priorities with a myriad of alternative allocations—for example, domestic education, health, income security, and foreign affairs. Within the constraint of limited resources, the decisive issue is always one of policy and related goals. In the American system of government, both foreign and domestic policy are the preserve of the civilian administration.

Defense decision-making, or military policy, then, cannot be considered independently. Fluctuations in defense spending are related to such exogenous factors as tax revenues, inflation, and the encumbent administration's views on balanced budgets. Further, the decisions of Republicans and Democrats regarding defense can be directed related to their positions on other issues.

Military policy can be usefully divided into (1) strategy decisions, and (2) structural decisions. Strategy decisions pertain to the size and use of force and include strength, composition, and readiness of forces. Such decisions as strategic and tactical deployments commonly embodied in war plans are also included. Strategy decisions are largely executive. In close consultation with his Secretary of Defense and Joint Chiefs of Staff (JCS), the President establishes high-level strategy. Structural decisions, on the other hand, pertain to the procurement, allocation, and organization of resources that implement the strategic units and require both executive and legislative action. The focus of structural decisions is the defense budget which is, in turn, part of the national budget and thus immersed in domestic politics.

Simply stated, defense decision-making is the conglomerate product of competing goals in which the relative weight given individual goals is dependent upon a highly unpredictable foreign environment and a fickle domestic environment.

Defense Decision-Making During the 1950's

The defense budget has always been the key to defense decision-making. It establishes the absolute magnitude of national resources that can be committed to security goals. During the 1950's, budgeting and defense planning were considered independently. Early in the budget cycle, the President provided guidance to the Secretary of Defense regarding a budget "ceiling" that he thought was economically and politically feasible for the next fiscal year. The Secretary of Defense then allocated a portion of this total to each service. The Services, in turn, suballocated their portions among their various programs. The Basic National Security Policy (BNSP) paper prepared by the National Security Council set guidelines on national strategy and priorities. Long-range defense planning for manpower and weapon systems was performed by the individual services based upon their estimates of the forces required to ensure our national security.

There was always a significant "gap" between the forces that individual services proposed were required to meet our national security objectives and those forces that they could actually procure. This was largely because little or no interservice coordination existed between defense plans. For example, prior to 1961, the airlift capability of the Air Force was not sufficient to transport the forces the Army was developing. The Army was planning forces and stockpiling inventory for a long conventional war, depending upon close-air support. The Air Force, on the other hand, was concentrating almost exclusively on aircraft for use in tactical nuclear war. Thus, even though the Air Force was committed to support the Army, divergent goals did not permit the Air Force to allocate sufficient resources to do so. The impact is self-evident; redundancy and imbalance seriously degraded military cost-effectiveness [3].

In addition, the basic framework of allocating a fixed budget by service, rather than by major mission (Strategic Nuclear Forces, Mobility Forces, Tactical Air Forces, etc.), complicated the task of achieving a balanced defense program. For example, each service made a contribution to the total military nuclear capability. The Army controlled the Minuteman missile system; the Air Force controlled an offensive missile system and bomber forces; and the Navy controlled the sea-based Polaris forces. Each service considered its program independently from the other services' programs; thus, nuclear strategy as a major mission was fragmented. Further, the Secretary of Defense received cost data by object classes—Procurement, Military Personnel, Installations, etc.—rather than by weapon systems—Strategic Nuclear Forces, General Purpose Forces, etc. This cost data was presented at the Department of Defense level on a year-at-a-time basis. Because inception costs of most programs are relatively small, many ultimately expensive programs were initiated with little hope of their completion at existing budget levels.

As the 1950's came to an end, our military posture actually included only the one option of nuclear deterrence. The capability of the Army to engage in an extensive limited war was highly questionable because of its dependence upon nonexistent strategic and tactical resources in the other services. In essence, the military effectiveness for tax dollar spent was seriously impaired by the management control system of the Department of Defense.

These problems, however, were neither unknown to nor accepted by the Eisenhower administration. Several attempts for their resolution resulted in a very favorable climate for reorganization by the Kennedy administration.

McNamara PPBS for Defense Decision-Making

By the late 1950's, the President, Congress, and many private citizens stressed the importance to national security that foreign, economic, and military policies be coordinated, and that imbalances in the force structure be eliminated. For example, the Rockefeller report, on the problems of the United States defense, recommended in 1958 that a start be made toward a budgetary system that "corresponds more closely to a strategic doctrine. It should not be too difficult, for example, to restate the presentation of the Service budgets, so that instead of the present categories of 'procurement,' 'military personnel,' etc., there would be a much better indication of how much goes, for example, to strategic air, to air defense, to antisubmarine warfare, and so forth." [4].

Other influential critics commented on the problems accruing from the planning and budgeting gap. General Maxwell Taylor stated: "The three Services develop their forces more or less in isolation from each other, so that a force category such as the strategic retaliatory force, which consists of contributions of both the Navy and the Air Force, is never viewed in the aggregate . . . In other words, we look at our forces horizontally when we think of combat functions but we view them vertically in developing the defense budget" [5].

The House Appropriations Committee, in 1959, expressed concern for the costly false starts plaguing research and development programs. They stated: "The system should recognize the necessity to eliminate alternatives at the time a decision is made for quantity production. It is this decision that is all-important. At this point there should be a full evaluation of (1) the military potential of the system in terms of need and time in relation to other developments, by all the military services, and (2) its follow-on expenditure impact if approved for quantity production" [6].

Finally, the analytical tools necessary for economic analysis of strategies and weapon systems were available in a usable form by 1961. In the late 1940's, Mr. Charles Hitch began to assemble the Economics Division at Project RAND. The group innovated and refined the application of quantitative economic analysis to the choice of strategies and weapon systems. This work is summarized by Hitch and Roland McKean in their book, "*The Economics of Defense in the Nuclear Age*."

Secretary McNamara enlisted the help of Hitch, from RAND, as his Comptroller, and Alain Enthoven, also from RAND, as Hitch's deputy for Systems Analysis. Together, they instigated the management philosophy commonly referred to as PPBS—Planning-Programming-Budgeting System. PPBS became the device through which centralized planning was accomplished. Through it, national security objectives were related to strategy, strategy to forces, forces to resources, and resources to costs.

In establishing the basis for PPBS, McNamara made a number of important reorganizations and changes. First, national security objectives were related to strategy through planning done by the Joint Chiefs of Staff (JCS). JCS, with tri-service representation, developed the basic planning document referred to as the Joint Strategic Objectives Plan, or the JSOP. It essentially projected a force structure. The force structure was stated in terms of major missions embodying all three services.

Secondly, cost-effectiveness studies were performed on the JSOP force structure. Economic, political, and technical considerations were interjected into the programming decisions resulting in the Five-Year Defense Plan (FYDP). These considerations were largely the product of McNamara's new staff aides referred to as systems analysts. The Systems Analysis group provided the means through which McNamara "short-circuited" the cumbersome bureaucracy of the Pentagon in effecting change. (What systems analysis meant to the Department of Defense and how McNamara used it will be described at a later point.)

The third change was initiated to inhibit beginning programs which were destined for abortion at later dates because of budget constraints. As mentioned earlier, when weapon system expenditures were viewed a year at a time, many programs would be started because of the relatively small resource commitment required during their research and development (R&D) phases. In order to limit such commitments, McNamara required that 10-year systems costs be developed in considering new programs. Ten-year systems costs included R&D, investment to equip forces with capability, and operating costs for 10 years. The timing and relative magnitudes of these costs are shown in Fig. 1.

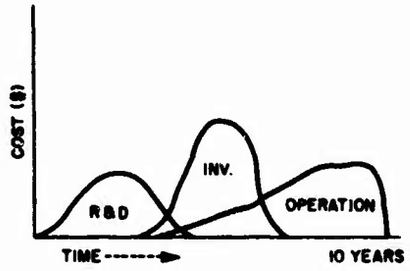


FIGURE 1. Weapons systems cost*

In considering weapon systems, discounted 10-year systems costs were used because a modern weapons system has a high probability of being obsolete in 10 years; and the relevant costs are related to keeping the system in a given state of readiness.

The fourth change consisted of a set of organizational alterations that were designed to better support PPBS. McNamara consolidated the supply and procurement systems into a DOD organization—The Defense Supply Agency (DSA). Also, he created the Defense Intelligence Agency (DIA) to provide relevant inputs into the JSOP planning process. Many other organizational changes were made that were centralizing in effect, but which also provided the necessary framework for decentralizing decision-making.

Office of the Assistant Secretary of Defense (Systems Analysis)†

From 1965 to 1969, the Systems Analysis staff was probably the most colorful and controversial group in modern government. Most popularly referred to as McNamara's "whiz-kids," the group has been characterized by being bright, but militarily inexperienced; skeptical of authority, but PhD conscious; esoteric, but iconoclastic; arrogant, but honest. In the past years, the staff developed a number of candid responses to critics of their studies who challenged their assumptions, but refused to provide any alternatives. Two such responses were: "It's better to be roughly right than exactly wrong," and "It's better to use bad data and good methodology than bad data and bad methodology." In any case, all of these characteristics probably do contribute to describing the profile of a systems analyst; however, a more accurate profile can be developed by describing the concept of systems analysis and how McNamara institutionalized it in the Department of Defense.‡

---

*Adapted from Charles J. Hitch, "Development and Salient Features of the Programming System," H. Rowan Gaither Lectures in Systems Science delivered at the University of California on 5–9 April 1965.

†In 1965, Alain Enthoven, the First Assistant Secretary of Defense (Systems Analysis) was appointed. Prior to 1965, Alain Enthoven was Deputy Assistant Secretary of Defense (Systems Analysis) to the Controller.

‡The approach is becoming widely applied in all aspects of the government. [illegible] Bureau Bulletin No. 66-3 requires department and agency heads to establish planning, programming, and budgeting systems.

While systems analysis has been described as "quantitative common sense" and the "general application of the scientific method," it has escaped precise definition. One of the reasons why is that, by its very nature, emphasis is placed upon the application of eclectic disciplines to a wide variety of problems. Concepts and techniques of systems analysis have been drawn from multiple disciplines, such as economics, mathematics, statistics, political science, and computer science; thus, it is difficult to align with one academic field.

A number of relatively simple principles have provided a basic framework which has been applied to most defense analyses in the past 9 years:

1. The data used in analysis must be verifiable, either by observation or deduction from plausible premises; the procedures employed in the analysis must conform to accepted rules of logic. Thus, the analysis is characteristically self-correcting.

2. Resources are always limited, but effectiveness is a function of creativity in organization.

3. All missions or activities can be accomplished in several alternative ways.

4. Alternatives should be compared by cost-effectiveness; more costly alternatives must have a commensurate increase in effectiveness.

Two curves provide the framework within which the systems analyst tries to place his analysis. The first curve is loosely called a cost-effectiveness curve (Fig. 2).

FIGURE 2. Cost-effectiveness curve

The cost-effectiveness curve, logically, illustrates the relationship between cost and effectiveness and diminishing marginal returns. That such a curve exists is central, and it is quite important where alternatives fall on the curve. To illustrate, consider the tons delivered into a contingency area during 30 days as a measure of effectiveness, and the number of aircraft and their support systems required to deliver the tons as dollar cost. The first squadron of aircraft and their support systems have a lower marginal productivity than the following squadron because of the initial setup cost of support systems such as air traffic control equipment, cargo-handling equipment, and maintenance resources. At some point, however, the curve turns sharply upward, and the increasing costs result in proportionately less and less effectiveness. This point may be reached when the preferred route becomes so saturated that no more aircraft are permitted to use the route. Additional aircraft are forced to fly alternate routes with longer "legs" resulting in lower payloads.

The second curve (Fig. 3) is loosely called the trade-off curve. The trade-off curve illustrates the concept of resource substitutions for accomplishing a mission. Any point on the curve represents a number of airplanes and ships that could accomplish a deployment mission. For example, p′ airplanes and q′ ships could accomplish the deployment mission, as could p airplanes and q ships. If the ratio of the distances a to b and b to c represents an equal cost ratio for airplanes and ships, the point e is the most cost effective number of airplanes and ships to accomplish the mission.

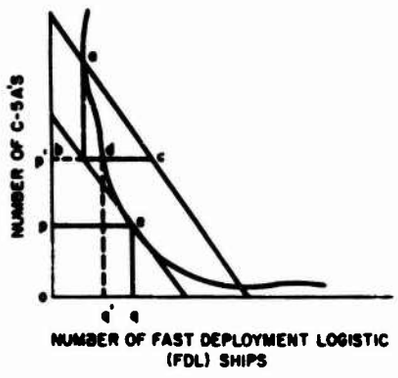


FIGURE 3. Trade-off curve

These curves represent a logic applicable to many problems of resource allocation. Quality and quantities can be traded off in a similar manner. Of course, the analysis is fraught with difficulties and complexities. Generally, the largest problem is measuring the multi-dimensioned concept of effectiveness.

Nevertheless, the function of the analyst is to draw out the cost and effectiveness of various alternatives so that the appropriate decision-maker can weigh the trade-offs and gain a better understanding of the relationship of costs and effectiveness. In the end, the defense decision-maker must exercise his own judgment as to whether the last increment of effectiveness (e.g., 3-day decrease in troop closure time with the enemy) is worth the cost of another increment of resources (e.g., an additional C–5A squadron).

Mr. McNamara's changes weren't so evident on the organization chart as they were on the locus of authority and the processes by which major decisions were made. McNamara found that bare military opinions were insufficient bases for making decisions. All too often, basic analysis principles were excluded from military studies. Thus, he insisted on seeing the data and reasoning behind recommendations. Although McNamara felt that no significant military problem could ever be wholly susceptible to purely quantitative analysis, he also felt that every aspect of the total problem that could be quantitatively analyzed removed one more element of uncertainty from the decision process [1]. Feeling most confident with studies which compared alternatives in terms of their costs and some solidly based criteria of effectiveness, he organized Systems Analysis to parallel the major defense missions. As experienced practitioners of the kinds of studies McNamara found useful, the systems analysts initiated, guided, and synthesized military research. Although their work sometimes competed with the work of the military advisers, Systems Analysis was designed to supplement the studies of the military advisers [5].

In order to forcibly impose a study discipline for decision-making on the military, McNamara delegated authority to Systems Analysis through the Draft Presidential Memorandum (DPM). DPM's consisted of 20 pages or less (excluding tables) and were the principal vehicles by which force-level* decisions were reached. The purpose of the memorandum was to study the force levels recommended in the JSOP, as well as alternatives. Using analytical tools, cost and objective achievement implications for feasible alternatives were subsequently set forth in the DPM. As previously exemplified, a

*Force levels are comprised of the resources required to satisfy an objective. In the JSOP and DPM, force levels may be expressed in units of aircraft squadrons, Air Force wings, Army divisions, missiles, ships, etc. The units also include personnel, equipment, and support resources required to make the unit operational.

3-day decrease in troop closure time, weighed against the cost of another C–5A squadron, could be one alternative offered. Reconciling the costs of various alternatives with the required force-level objective was the task of the ultimate decision-maker. DPM's were decision documents for the President. Conversely, Defense Guidance Memorandums (DGM) were transmitted to the Secretary of Defense for decisions. Other than this one difference, the two documents were the same. Together, they provided the basis for changing the Five-Year-Defense Plan (FYDP)—the basic planning document. McNamara's DPM's and DGM's covered the 20 functional areas listed in Table 1.

The responsibility for a DPM was assigned to a systems analyst. He then accumulated data and performed and coordinated analysis leading to a basis for decisions by the Secretary of Defense or the President. Although the analysis cycle was continuous, it is useful to think of the JSOP as the first major document starting a new cycle. Figure 4 shows the process.

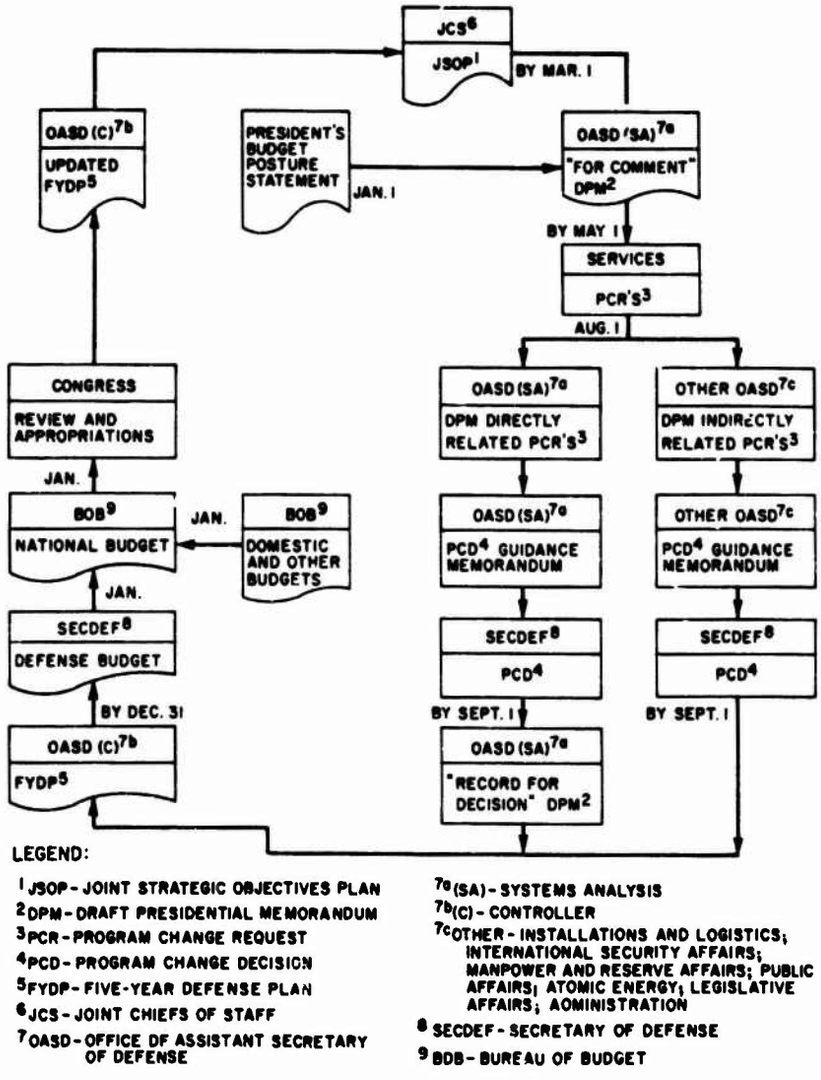


FIGURE 4. McNamara Planning-Programming-Budgeting System (PPBS) Cycle

The JSOP, along with the President's Budget Posture Statement, established the basic military strategy for the DPM. Rarely, if ever, did the DPM author look to the Services for unilateral contri-

butions to strategy. As a beginning point for analysis, the DPM author used the previous year's FYDP and "Record of Decision" version of the DPM. From this base, he conducted discussions with the Services, JCS, and other members of the Office of the Secretary of Defense (OSD) staff in order to acquire data and rationale to support the development of the next DPM. Additionally, the author examined relevant studies and analyses performed by the Services and other agencies. The synthesis and integration of the author's own analyses culminated in the publication of the "for comment" version of the DPM. The "for comment" version triggered force programming.

TABLE 1. *Draft Presidential Memorandums/Defense Guidance Memorandums*
(Presented in the sequence in which normally prepared)

DRAFT PRESIDENTIAL MEMORANDUMS (DPM's)
Logistic Guidance for General Purpose Forces
Asia Strategy and Force Structure
NATO Strategy and Force Structure
General Purpose Forces
Land Forces
Tactical Air Forces
Anti-Submarine Warfare Forces
Escort Ship Forces
Amphibious Forces
Naval Replenishment and Support Forces
Mobility Forces
Strategic Offensive and Defensive Forces
Theater Nuclear Forces
Nuclear Weapons and Materials Requirements
Research and Development
Military Assistance Program
DEFENSE GUIDANCE MEMORANDUMS (DGM's)
Indirect Support Aircraft
Pilot and Navigator Requirements, Inventories, and Training
Manpower
Shipbuilding

The May 1 "for comment" version was submitted to the Service Secretaries and JCS for "line in, line out"* changes. Within 4 weeks, the Services submitted to Systems Analysis their comments and rationale along with their Program Change Requests (PCR). August 1 was the deadline for submitting all PCR's. The DPM author then prepared a Program Change Decision (PCD) Guidance Memorandum summarizing the DPM position and the Services' positions, presented a brief evaluation of the issues and alternatives available, and made a recommendation to the Secretary of Defense. A complete set of the JCS's and Services' comments was attached to the Guidance Memorandum to ensure that the comments were not distorted in the process. The Secretary of Defense examined the PCD Guidance Memorandum, requested amplification if required, and issued guidelines for preparing the PCD. Based upon the guidelines, the DPM author prepared the PCD, coordinated it with the JCS and Services, and forwarded it along with any comments to the Secretary of Defense. The Secretary considered any

*"Line in, line out" changes refers to the process of crossing out words or lines in an original document so that the words are still legible and designating revisions by underlining.

comments of the JCS or Services, reached a decision, and approved the publication of the PCD. Once the Secretary signed the PCD, it was forwarded to the OASD (Comptroller) for budgetary action.

After publication of the last PCD (about September 1), until November 1, key issues raised by the process were further debated and negotiated; and, during this period, supplemental decisions could be made. Also during this time a "Record of Decision" DPM, incorporating all changes since the PCD, was prepared and issued for each DPM. These "Record of Decision" DPM's were then used to again update the FYDP and support the President's defense budget submission to Congress in January.

Since Systems Analysis controlled the DPM, the basic force programming document, and supplemental documents required to alter the FYDP, the group exercised a great deal of power in influencing defense decisions. It is generally agreed (although controversy always exists) that the result has been a substantial rise in the quality of research and, ultimately, a higher regard for military advice than at any time in the relatively brief history of the Department of Defense. Cost-effectiveness studies have tended to clarify which issues are best left to military judgment.*

With the improvements, however, have come problems. For example, a precise definition of objectives is imperative to effective systems analysis. During the Kennedy-Johnson administrations, no formal cabinet body existed to establish and state national security objectives. Instead, the "threat" to national security was estimated through intelligence appraisals derived from both the military and the Central Intelligence Agency. The JCS then developed the Joint Strategic Objectives Plan (JSOP) which included a recommended force structure to meet the estimated "threat." In lieu of a body which formally stated objectives then, the JSOP became the Department of Defense document which performed that function.

At times, aggressive systems analysts, for the sake of effective analysis, imputed objectives where those available in the JSOP were poorly defined. Once clarified, the analyst's objectives often gained general acceptance. As an example, the size of the conventional Army, Navy, and Air Force was based upon an accepted defense objective of maintaining the capability to fight simultaneously a land war in Europe and Asia, plus a minor conflict in the western hemisphere. The origin of the "two majors and a minor contingency simultaneously" is a controversial subject. Nevertheless, one of its first appearances was in Systems Analysis where the scenario was designed as a "worse case criterion" for measuring the capability of airlift and sealift resources to deploy forces.

At other times, systems analysts have indiscriminately imposed esoteric analyses upon the Services. Some military officers, feeling that they lost status, resented what they regarded as a failure to recognize their contributions. In some cases, even though the analytical work of the military staffs improved dramatically, it may not have received due consideration and credit.

Regardless of the sources of these animosities between Systems Analysis and the military, frictions exist within the Department of Defense which endanger continuance of the Systems Analysis office. Administratively, eliminating the Systems Analysis office has some advantages. The office has been stigmatized; and, along with avid supporters, it has acquired radical critics in Congress and the Pentagon. The emotions triggered by the "Systems Analysis whiz-kids" title obviously inhibits its flexibility in adapting to a relevant role.

In addition to the political biases afflicting the Systems Analysis office, its basic mechanism of influence, the DPM process, also has intrinsic shortcomings. During the Eisenhower administration,

*Ironically, while credibility of subjective military judgment has increased largely due to Systems Analysis, the credibility of Systems Analysis studies seems to have decreased due to their failure to take into account subjective factors.

control of defense procurements was through budget ceilings; while during the Kennedy-Johnson administrations, control was maintained through force-level ceilings as expressed in the DPM. This change of control deemphasized the defense budget as a constraint. According to the Kennedy-Johnson administrations, "The country can 'afford' to spend as much as necessary for defense" [2]. Thus, limits for military spending were expressed primarily in terms of force size, and secondarily in terms of dollars. However, the exact force size necessary to meet national security objectives involves a great deal of conjecture and uncertainty. Because of differing points of view, the systems analysts and the military seldom agreed in their estimations of the forces needed to meet an enemy threat. As a general rule, the Systems Analysis group tended to estimate needs more conservatively.

These differences in judgment led to a perennial tug-of-war throughout the budget cycle. Because most disputes involved force size (e.g., number of wings, divisions, etc.), the Services tried to incorporate as much as possible in their weapon systems within the limits which they view as "fixed force ceilings." For example, although only one aircraft or ship may be recommended by a service and approved by the Secretary of Defense, this one piece of equipment may have been subsequently "goldplated" to include multipurpose features. To illustrate, the mere avionics of an F-4 fighter cost considerably more than a total F-100 fighter did in 1961. Obviously, many technological and economic factors account for the increased cost of a fighter. Nevertheless, an element of "goldplating" must be suspected.

This practice has ultimately meant spiralling costs for the Defense Department and unjustified requests for increased capabilities, regardless of expense. There has been little incentive for the Services to stay within a budget ceiling, because they realize that such goldplating will probably not affect their other programs as it would have in earlier years when they operated within a fixed budget. The extra costs incurred may well have come from an add-on to the total defense budget or have been siphoned from the other Services' programs.

A second problem has been that the military tended to request everything in the hope that something would slip through Systems Analysis. Centralized analysis could not be possibly used to evaluate each of the proposals objectively. Thus, systems analysts tended to sort through the barrage of proposals by performing analysis which roughly supported negotiation positions for the Secretary of Defense. This is precisely the area in which Systems Analysis has been indicted for taking the dominant role in the weapon system selection decision process.*

Laird/Packard PPBS for Defense Decision-Making

Probably due primarily to the difficulties involved in the transition of administrations, the 1969 calendar year budget cycle was executed through the McNamara DPM process with the exception of a few minor changes (See Fig. 5). The number of DPM's was reduced to two. In addition, eight Major Program Memorandums (MPM's) were introduced for annual major programming issues decided by the Secretary of Defense. MPM's replaced and consolidated many previous DPM's. Two DGM's were developed for nonrecurring major issues decided by the Secretary of Defense. Table 2 lists the 1969 revised DPM's, MPM's, and DGM's. A notable difference from the previous year's process, however, was that the FYD. was not updated for "out years" (i.e., years beyond Fiscal Year 1971).

*With the departure of both McNamara and Enthoven, the Services "dug up" buried proposals, such as manned bombers, quiet submarines, and new missiles to resubmit to the administration. Because of the many uncertainties involved, the success of "objectively" discounting the proposals with trade-off and cost-effectiveness analyses that have already been performed is small. If the proposals should be reevaluated using this technique, a high probability exists for starting some programs that must ultimately be cancelled because of budget constraints, and also, the risk of unbalanced force structure increases.

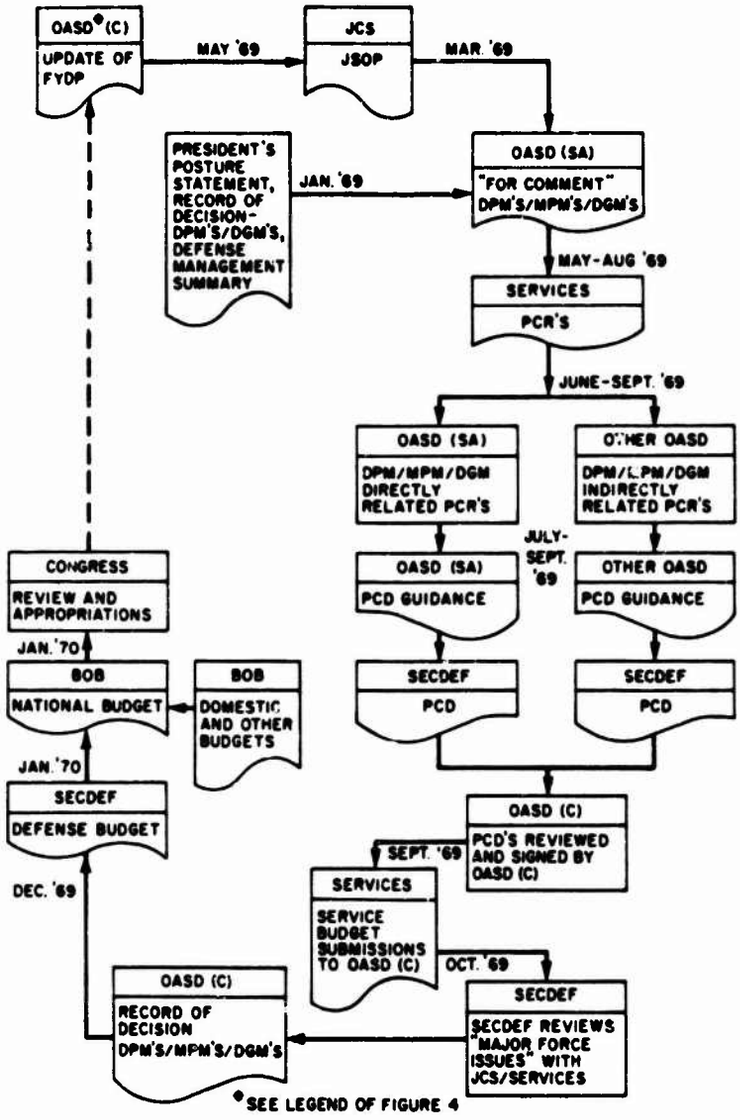


FIGURE 5. Calendar Year 1969 Planning-Programming-Budgeting System (PPBS) Cycle

As pre-Nixon people left during early 1969, the Systems Analysis office was slightly reorganized. On January 31, 1969, one of the original McNamara "whiz-kids" was appointed *Acting* Assistant Secretary of Defense (ASD) for Systems Analysis.

Beginning in early summer and before the Fiscal Year 1971 budget had been submitted to Congress, the Laird/Packard PPBS began to take form. The theme is decentralized decision-making. The reduced role for the Systems Analysis office was also correspondingly clear. On December 11, 1969, the Acting ASD for Systems Analysis (Dr. Ivan Selin) submitted his letter of resignation citing the fact that it had become clear that the Senate would not confirm his position.* Less than a week

*In response to the letter of resignation, Laird, in part, wrote "Unfortunately, a number of people in various pursuits—in Congress, in the Executive Branch, and from outside the Government—have misunderstood the role of Systems Analysis. This misunderstanding has, in all candor, been translated to a mistrust of the key officials in the Systems Analysis office. The mistrust, ironically, has been exacerbated by the fact that you and your staff have been so effective in discharging your assigned roles."

TABLE 2. *Draft Presidential Memorandums, Major Program Memorandums, and Defense Guidance Memorandums*

DRAFT PRESIDENTIAL MEMORANDUMS (DPM's)
General Purpose Forces Strategic Forces
MAJOR PROGRAM MEMORANDUMS (MPM's)
Land Forces Tactical Air Forces Naval Forces Amphibious Ship Forces Mobility Forces Theater Nuclear Forces Manpower Research and Development
DEFENSE GUIDANCE MEMORANDUMS (DGM's)
Logistics Nuclear Stockpile and Materials

later, the President sent a nomination to the Senate for a replacement and it was immediately confirmed.* Since then, many of the Directorate positions in Systems Analysis have been staffed with military leadership. Replacing civilian leadership with military leadership weakens the impartiality of a central power by introducing the dysfunction of "vested interests." The Systems Analysis Directorates are faced with many decisions n which the best course of action for the Secretary of Defense violates the interest of a particular military Service. The existence of the inherent goal incongruence intimates objectivity, and thereby, the credibility of decisions. The probable effect is that few issues will be adjudicated by Systems Analysis.

Figure 6 shows the sequence of events for the Laird/Packard PPBS. One of the strong points of the system is formal goal setting by the revitalized National Security Council. A second strong point is the concept of "fiscal guidance" which communicates to the Services the hard realities of political considerations and budget ceilings. Control of budget ceilings is the main Office of Secretary of Defense (OSD) management control mechanism. The decentralization of force level and mix decision-making to the Service Secretaries is real. The Program Objectives Memorandum (POM) is the central document in the PPBS replacing the DPM. It is prepared by the Secretaries of the Military Departments and embodies the total program requirements by major mission and support categories necessary to support their assigned missions. OSD will check the POM's for adherence to fiscal guidance and summarize them into a Program Decision Memorandum (PDM). In turn, the PDM is proposed to be used for updating the FYDP.

The planning process seems to be the weak link in the new PPBS. OSD apparently has no effective control device to ensure realistic planning by the Services. If the past is any indicator, the Services will be quite optimistic concerning total weapon systems cost. As a result, "out-years" planning will be overly optimistic and may result in aborted development programs in order to stay within budget ceilings.

Moving from design to organization for PPBS, a real question is whether the Services have the analytical resources to support decentralized force level size and mix decision-making. The com-

*The nominee was Dr. Gardiner L. Tucker, Principal Deputy Director, Defense Research and Engineering.

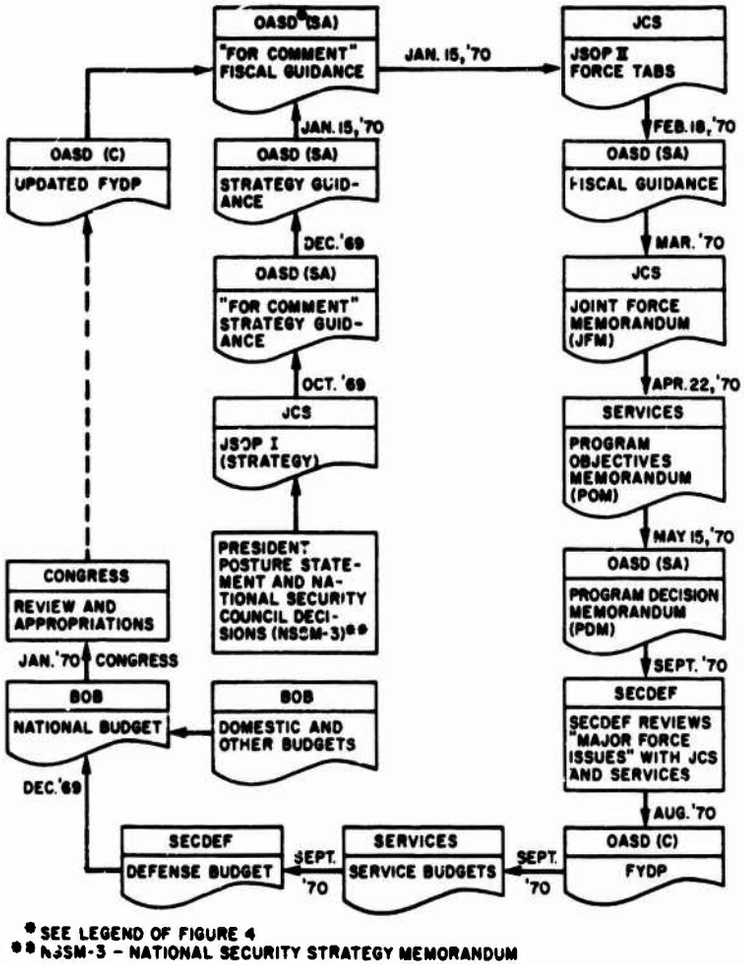


FIGURE 6. Calendar Year 1970 Planning-Programming-Budgeting System (PPBS) Cycle

plexity of the decisions requires systems analysis at its best. On the optimistic side, there is some evidence that in the past years the Systems Analysis office has forced analytical parity onto the Services. On the pessimistic side, analytical resources are especially scarce. Recruitment problems have been aggravated for the Services by their recent image disadvantage associated with Southeast Asian involvement.

Unfortunately, the actual effects of a PPBS can be only assessed in the long run. Effective long-range planning is the central issue. Formal mechanisms to guide and control planning are essential, and the new PPBS seems weak in formal planning mechanisms for maintaining a balanced force structure. The Secretary of Defense has indicated that he is going to hold the Service Secretaries unequivocally accountable for their programs; however, by what standards or criteria this will be achieved is unclear. Communicated and accepted standards and measures of performance are basic to effective management control.

On balance, the Laird/Packard PPBS has integrated many of the successful defense management tools of both the Eisenhower and McNamara systems: formalized objectives, fiscal guidance, costs by major programs, and systems analysis. The major change from the previous system is decentralization. It is a well-accepted management principle that, in order to work, decentralization must be real.

The decentralization under the new PPBS is real. Nevertheless, the analogy persists of the ill-fated company which scraps its manual payroll system for an untested computerized system. The old centralized PPBS has been scrapped for the new decentralized PPBS. Presently, the risk is high; but if the debugging process can be tolerated, the system may prove many times better than its predecessor.

REFERENCES

[1] Kaufman, William W., *The McNamara Strategy* (Harper and Row, New York, 1964), p. 295.

[2] McNamara, Robert S., "Managing the Department of Defense," Civil Service Journal **4**, 1–5 (1964).

[3] McNamara, Robert S., *The Essence of Security: Reflections in Office* (Harper and Row, New York, 1968).

[4] Rockefeller Brothers Fund, *International Security the Military Aspect*, report of Panel II of the Special Studies Project (Doubleday, Garden City, N.Y., 1958), pp. 58–59.

[5] Taylor, Maxwell D., *The Uncertain Trumpet* (Harper and Row, New York, 1959), p. 123.

[6] U.S. Congress, Committee on Appropriations, House Report No. 1561, *Report on Department of Defense Appropriations Bill*, 1961 86th Congress, 2d Session (Government Printing Office, Washington, 1960), p. 25.

THE FAST DEPLOYMENT LOGISTIC SHIP PROJECT: ECONOMIC DESIGN AND DECISION TECHNIQUE

David Sternlight

Litton Industries
[illegible] Hills, California

ABSTRACT

This paper describes the way in which economic analyses, particularly life-cycle cost analyses and tradeoffs were structured for use as an integrated analysis and design technique at all levels of the Contract Definition of the Fast Deployment Logistic Ship. It describes system, subsystem and major component economic analysis and design methodology as well as economic analyses of special subjects such as the ship production facility design. Illustrations are provided of several major system parametric studies and of shipyard and manning/automation analyses.

I. INTRODUCTION

The purpose of this paper is to describe the application of economic analysis, particularly life-cycle cost analysis, to the Contract Definition design of the Fast Deployment Logistic Ship system, subsystems and components. Overall performance and mission envelopes were specified by the Navy for this, the sea-lift portion of the U.S. Strategic Rapid Deployment System. A production schedule that could not be met by any existing shipyard was required, and it was made clear that contractors were expected to design a highly modernized or completely new facility, heavily mechanized to reflect design consistent with the best modern shipyards of Europe and Japan. The purposes of the competition were described by the Navy as three-fold:

1. To design and develop a high-performance rapid response ship capable of carrying infantry division cargo for up to 3 years under conditions of controlled temperature and humidity and able to respond rapidly to an emergency in major areas of the world, delivering its cargo rapidly in ports or over unimproved beaches in order to mate with airlifted troops.

2. To introduce systems analyses, life-cycle cost analysis, and the Contract Definition process into the design of Naval ships.

3. To make a trial application of the total package approach for ship procurement.

From these requirements, and performance and mission envelopes, a Contract Definition analysis and design of the ship was conducted by three major competitors. Cost and benefit analysis was performed at every stage of design from the system conceptual phase through facility and production planning. Life-cycle cost analysis was not only a formal program requirement, but a major evaluation criterion. Therefore, it was necessary to plan the Contract Definition Phase and to design techniques for economic analysis of overall hardware characteristics, production facility location, production facility design, and integrated logistics support systems, as well as for such analysis in the detailed engineering decision process leading to physical and performance parameters of the system, subsystems, and components.

The evaluation criteria for the FDL Contract Definition product included technical content of ship design, military effectiveness, and life-cycle cost. Military effectiveness was fully defined through

the specification of a figure-of-merit, and a systems analysis problem. Ship and system parameters in the systems analysis problem were to be determined to minimize system life-cycle cost subject to side constraints on fleet delivery capacity and delivery time. After establishing certain key ship and system parameters, the main quantitative analytic criterion became minimum life-cycle costs subject to side constraints expressed as performance and mission envelopes. A speed envelope, for example, was specified. Within the overall decision rule of minimum life-cycle costs, three classes of analyses were performed. System parametric studies established fleet and ship characteristics to satisfy performance and mission requirements and the systems analysis problem with a high figure-of-merit and low life-cycle costs. Through appropriate analytic sequencing, those parameters which were related to cost effectiveness were first explored and their values established. Subsequent analyses could then be performed using a minimum life-cycle cost decision rule. Engineering economists performed special studies of such subjects as production facility site selection, internal production facility configuration, and manning/automation. Although analytic methodology was hand tailored to each problem, the structure within which these analyses were conducted was the life-cycle cost structures established for the entire program. Many extensive hardware life-cycle cost tradeoffs were also conducted, using a standard analytic method an a prescribed series of "object-related" cost categories. A managerial technique was developed to permit a modified form of subproject organization to overlay the functional organization of the Litton Contract Definition team. Hardware subsystems analysis was performed by a number of joint teams, each including a subsystem engineering design expert, a life-cycle cost analyst, a reliability and maintainability analyst, a human factors analyst, and an integrated logistic support specialist. In this way, subsystems were designed to achieve the benefits of reliability, maintainability, and effective integrated logistic support analysis within the framework of joint minimization of total subsystem life-cycle costs within effectiveness envelopes. Tradeoffs between initial investment costs, direct operating costs, manning costs and maintenance and repair costs for differing levels of reliability and different maintainability configurations were an integral part of the overall subsystem design process. Finally, a format for life-cycle cost analysis in the selection of components was developed to permit engineering specialists to configure components of subsystems for minimum total life-cycle costs.

The common thread in all these analyses is the tool of discounted present value cash-flow analysis often used for the comparison of capital investment alternatives. In this case, all flows were considered: direct and indirect government and contractor investment costs including hardware construction, systems management, systems evaluation, training, data, industrial and operational facilities, initial spares and repair parts; and operating and support costs including manning, direct operations, maintenance and repair, material, and indirect operating support. An integrated engineering design model was developed and programmed for the efficient parametric analysis and tradeoff of many thousands of different system hardware configurations. The model included an engineering design optimization portion and a life-cycle cost portion. For each set of hardware parameters a most efficient hardware configuration was selected and its life-cycle costs determined. Many hundreds of these "most efficient" hardware configurations for varying parameter sets were compared before the final systems hardware configuration was selected. For the analysis of subsystems and components, tradeoffs were performed in detail by the teams already described, using an overall system model when the costs of other portions of the system were affected by the selection of particular subsystem or component alternatives.

As a result of the complete, coherent application of life-cycle cost analysis as an engineering decision-making tool from system to component, a step by step economic justification of the entire

system and the rationale for its selection exists. It is possible to see how decisions at any stage affect and are affected by previous and subsequent decisions. It is also possible to explore the decision chain when changes to the system are contemplated in order to provide an efficient method for the analysis of the economic effect of these changes.

II. LIFE-CYCLE COST ANALYSIS AND INTEGRATED SHIP AND SYSTEM DESIGN

The step-wise economic analysis performed (Fig. 1) in order to design a ship and system at all levels

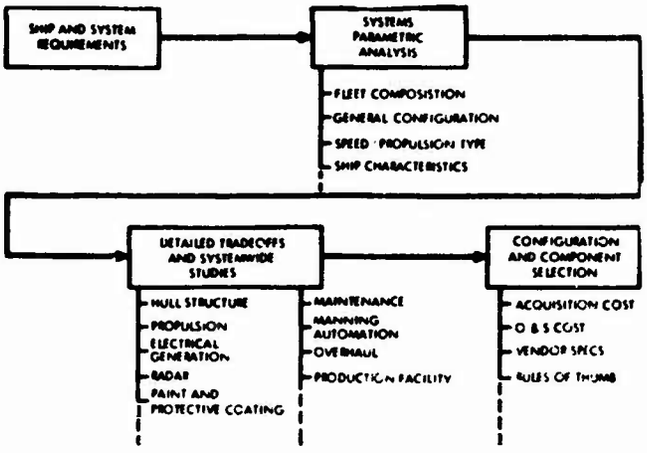


FIGURE 1. Stepwise economic analysis

while maximizing figure-of-merit, minimizing life-cycle cost, or achieving both objectives, began with the determination of ship and system requirements. The Navy specified a series of performance and mission envelopes which defined the ranges within which certain critical ship design parameters must fall. They specified a systems analysis problem which was in the form of a heavily parameterized resource allocation problem. The major mission of the FDL ship is to deliver infantry division force cargo in response to an emergency, to specified destinations in specified amounts. The systems analysis problem defined the possible origins for the FDL fleet, the amounts of cargo prepositioned at various points, the ship loading conditions prior to the initiation of an emergency deployment, and the cargo amounts, delivery destinations, and delivery times to meet the military requirement. The problem did not specify the speed, cargo capacity, or other ship characteristics. These parameters had to be determined through exercising the systems analysis problem, to meet the delivery time and cargo capacity requirements at lowest life-cycle cost. This implied the choice of ship size, ship speed, fleet size, and ship prelocation. A number of side requirements (such as ability to transit the Panama Canal) were included which provided additional constraints on the ship and fleet design parameters. At the systems level, then, our objective was to define fleet composition, general ship configuration, speed and propulsion type, and detailed parametric characteristics of each ship to satisfy the performance envelopes, the side constraints, and to maximize the figure-of-merit specified by the Navy. The sequencing of the analysis (Fig. 2) shows the process of figure-of-merit maximization.

The problem was to maximize the classical "transportation momentum" measure:

$$\frac{\text{Speed} \times \text{Capacity}}{\text{25-year discounted life-cycle cost}}.$$

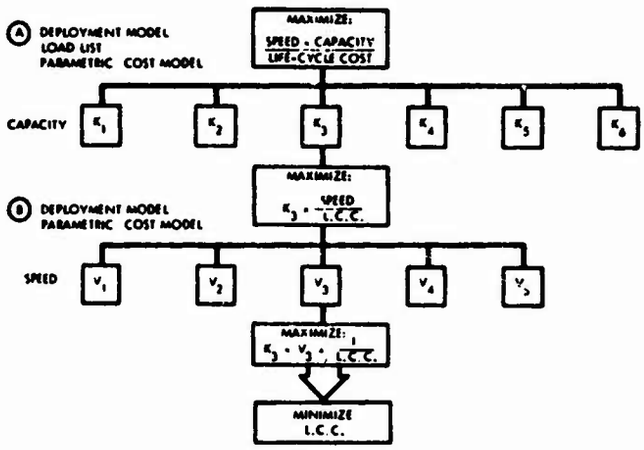


FIGURE 2. Figure-of-merit analysis

As a first step, consider the determination of individual ship capacity. Through the use of the deployment model, the use of a detailed load list, and the use of a parametric engineering design and life-cycle cost model, various alternative capacities and hence fleet sizes, were determined in order to maximize the figure-of-merit, by varying individual ship capacities subject to a fixed total fleet capacity. These analyses made it clear that a particular fleet size and ship capacity resulted in least life-cycle costs and maximum figure-of-merit over all speeds in the range of interest. With the capacity determined, the next step was to maximize the speed-cost ratio subject to the other performance and mission envelopes. Here again, the deployment model and the parametric engineering design/life-cycle cost model were used to perform analyses at many different speeds. It became clear that the systems analysis problem would be satisfied by a range of speeds within the speed envelope, and that a minimization of life-cycle costs for speed, with due regard to design risk, would also minimize life-cycle costs in the systems analysis problem. A speed was thus determined resulting in lowest life-cycle costs, considering design risk. With the speed and capacity fixed, the figure-of-merit became: maximize K/LCC with K constant, which is equivalent to minimizing life-cycle costs. Our subsequent analysis and design could be conducted, within the fleet and ship characteristics already specified, with the objective of minimizing life-cycle costs subject to remaining mission and performance requirements. A parametric analysis of life-cycle costs and figure-of-merit for different speeds and power plants (Fig. 3) shows that for the four major types of power plants considered at the systems level, Type I clearly has lower life-cycle costs and a higher figure-of-merit at any speed.

Power Plant Type I, therefore, was dominant within the range of speeds considered for this problem and was selected. Having selected Power Plant I, further analysis indicated that the lower the speed, the lower the life-cycle costs. At this point, a selection of speed was made based on the findings of this analysis together with due regard for design risk. With the fleet composition, general configuration, speed and propulsion type determined, the next step was to specify ship parameters. Physical parameters of a ship, such as beam, length, block coefficient, and the related endurance and stability characteristics for a ship of a given speed and payload are closely interrelated. One cannot consider curves of life-cycle cost versus ship length without due regard for the variation in other parameters. Many ships of the same length, but with different beams and block coefficients will carry the specified cargo. We see (Fig. 4) many such ships plotted against the figure-of-merit which, at this point, is equivalent to the inverse of life-cycle costs. The intersections represent physically realizable ships. As the beam

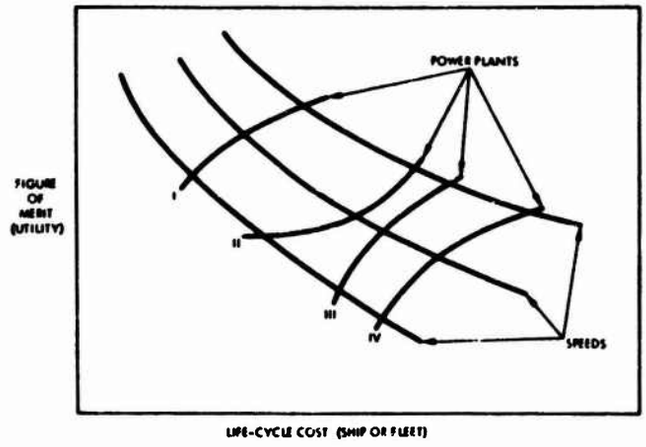


FIGURE 3. Speed/propulsion economic analysis

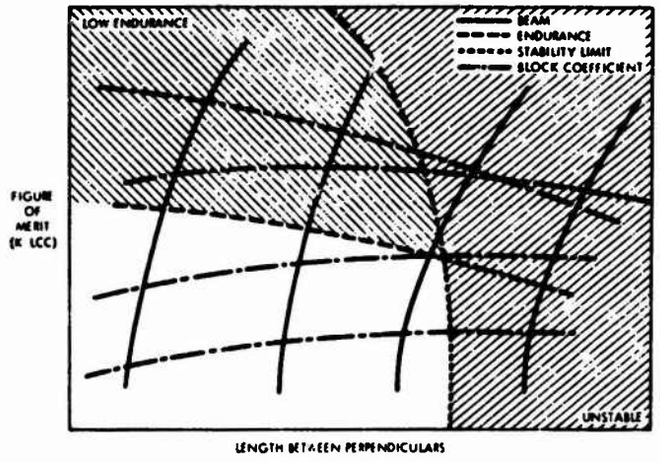


FIGURE 4. Ship characteristics analysis

decreases we reach a region of instability and ships to the right of the stability limit are unacceptable. Another side constraint, the endurance limit, is shown as a dashed line. Ships of low endurance do not meet mission requirements. The set of acceptable, physically realizable ships, forms a small subset of all possible ships having acceptable characteristics and meeting the payload requirement. Through the use of many such analyses we determined the ship characteristics.

III. LIFE-CYCLE COST STRUCTURE AND MODELLING TECHNIQUES

The first step in developing a coordinated approach to life-cycle cost analysis is to define the cost variables of interest. The first step in doing this is to define the basic ground rules for life-cycle cost analysis. A key expression of the basic ground rules is to consider all costs which occur on account of the system of interest, while ignoring costs that would occur whether the system existed or not. Given these ground rules for assessing the applicability of particular costs to the program, the next step is to develop a life-cycle cost structure (Fig. 5). In this structure, system costs are divided into the three main phases of the life of the system: development, acquisition, and operations and support. These costs are further broken down: acquisition into contractor and government costs; contractor

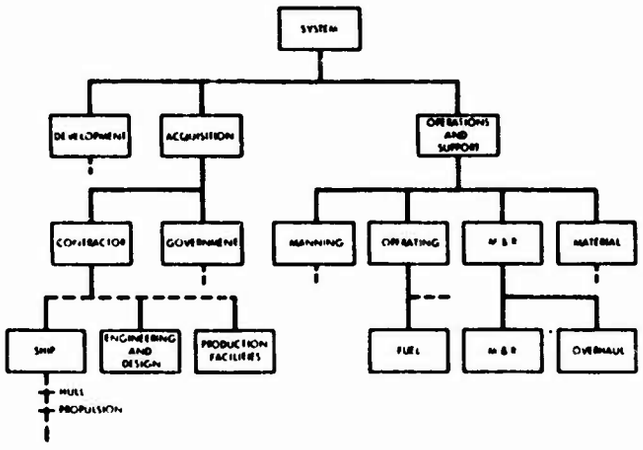


FIGURE 5. Life-cycle cost structure

costs into ship construction costs, engineering and design costs, production and facilities costs, management and technical costs, initial spare parts costs, and many other elements. Government costs are similarly broken down into appropriately detailed elements. The operations and support phase of the systems life is broken down by major resource categories used during this phase. These categories include manning, direct operating costs, maintenance and repair and related costs, materials costs, administrative costs, and other major categories. These costs are further broken down into appropriate subcategories such as fuel, maintenance and repair and overhaul. The basic structure for the FDL system was developed by the Navy; contractors elaborated the structure at the finer levels of detail. This permitted the comparison of competing contractor's costs using a common basic structure related to the way in which historical data on systems costs have been collected in the past. This structure is the key to all life-cycle cost analyses: system level analyses, subsystem analyses, and detailed engineering design analyses. All of these analyses involve the balancing of different elements of the overall cost structure against each other. For example, to evaluate equipment reliability, if two alternative equipments are available both meeting the minimum reliability requirements for the mission, one can determine whether the higher reliability item is justified by conducting a life-cycle cost tradeoff. The cost elements for equipment acquisition and initial spare parts are balanced against the operations and support costs over the life of the system for maintenance and repair, overhaul, and repair and spare parts. Instead of a series of such tradeoffs, the overall subsystem life-cycle cost tradeoff is used, simultaneously balancing reliability factors, training factors, manning and automation factors and many others. The cost impact of these diverse variables is assessed in the life-cycle cost tradeoff of the different subsystem design alternatives meeting the non-cost mission and performance requirements. The basic process of using the life-cycle cost structure to simultaneously balance many costs runs through our entire analytic process.

Having defined the life-cycle cost structure, the next step is to develop cost estimating relationships. These cost estimating relationships are of two major kinds. The first is an accounting relationship which indicates the structural breakdown of life-cycle costs. It describes those elements which are totals of lower level elements in the cost structure so that all summary elements (mechanical totals) are properly identified. Another kind of cost relationship is the parametric cost estimating relationship (Fig. 6), which describes the relationship between elements of cost and of physical performance, systems environment, and historical behavior.

① PROPULSION INSTALLATION LABOR COST $_G$

$$A \cdot B(SHP_G)^C$$

A : COST FACTOR PER MAN HOUR
B,C : HISTORICAL DATA REGRESSION COEFFICIENTS (ADJUSTED)
SHP : DESIGN HORSEPOWER, PLANT TYPE G

---

② FUEL COST $_{GH}$

$$\sum_{I=1}^{SHIPS}\left(\sum_{J=1}^{YRS}\left(\sum_{K=1}^{MODES}\left(\sum_{L=1}^{KNOTS} SFC_{GHL} \cdot SHP_{GL} \cdot T_{JKL} \cdot C_H\right)\right)\right)$$

SFC = SPECIFIC FUEL CONSUMPTION, PLANT TYPE G, FUEL H, SPEED L
SHP = HORSEPOWER, PLANT TYPE G, SPEED L
T = TIME AT SPEED L, MODE K, YEAR J
C = COST OF FUEL H

FIGURE 6. Cost estimating relationships

The first cost estimating relationship, illustrated for installation labor costs for a particular propulsion plant type, is derived through stepwise linear regression of historical data. A large number of regression analyses were conducted of historical data on ship materials and labor costs. Many different structural relationships were examined in this statistical cost analysis. The quality of each of these regressions was evaluated using multiple correlation coefficients, coefficients of variation, root mean square error, Durbin-Watson statistic, Theil U-statistics, and other measures. Statistical cost estimating relationships were thus structured and parameterized for the hardware costs associated with the ship. The example illustrated shows an exponential relationship which has proved extremely useful in practice for a variety of situations. The cost of installation is expressed for a "first ship" as a function of the cost per manhour and a historical function of shaft horsepower. The historical data used are adjusted to a constant dollar base to make costs in different years comparable. Overhead cost equations relate material and labor costs in the model, and appropriate learning curve computations are performed to develop the details of the ship acquisition cost contribution to the total life-cycle in the model.

The second type of parametric relationship, illustrated for fuel costs, is a cost estimating relationship based on engineering data and computations and descriptions of the environment in which the ship must operate. We see two engineering factors: the specific fuel consumption (SFC) based on a family of curves at different horsepowers for various propulsion types using specified fuels, which is derived from analytic and measurement data relating to these plants, and shaft horsepower (SHP) for each plant, derived from detailed analysis of the physical configuration of the ship in question, a large body of empirical ship resistance data, and information about the plant type and the plant weight, fuel weight, and other ship weights. These SHP curves summarize the horsepower required to drive the ship at any particular speed. A steaming profile, specified by the Navy, is used to indicate the various modes of operation, the times during which the ship will operate in these modes, and the percentage of total time in each mode spent at each speed. Combining the above factors with cost of fuel, we derive the annual fuel cost for any given plant type using any appropriate fuel, also considering the fleet size and the number of years of ship operation.

Having derived the cost estimating relationships for the model, the next step is to combine them in appropriate sequence (Fig. 7) in order to compute the life-cycle costs of the entire system. This sequence is a function of the relations between elements and their subtotals and totals, of the phasing of the program, and of the parametric relationship between elements. Investment costs for spare parts,

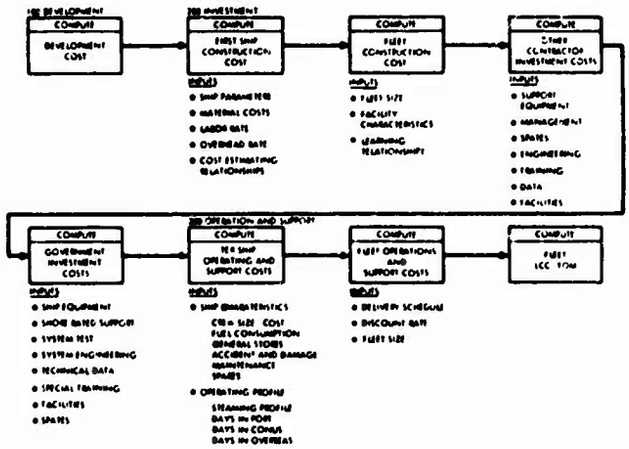


FIGURE 7. Life-cycle cost analysis model flow

for example, are related to ship parameters and construction costs by subsystem. Operations and support costs are related to the operational doctrine of the ship, its parameters and subsystem costs. Fleet costs are based on per ship costs and fleet-related factors not assignable on a per ship basis. For example, the creation of a management organization to supervise the construction and operation of the ships will require an initial infrastructure before the first ship is delivered. As the ships are delivered, additional personnel will be added to the management structure. Many cost elements contain such fixed and variable portions. In the model (Fig. 7) development costs, while sunk, are shown for completeness. These costs and some others do not vary with the ship design changes during Contract Definition nor do they affect the outcome of any tradeoffs.

Such a model could be used during concept formulation; many more elements would then be variable. When the fleet is operational, on the other hand, the life-cycle cost model will contain many more fixed elements. Toward the end of the life of the system, the life-cycle cost model would evolve into a historical data base for the program rather than a collection of variable relationships.

Investment costs are computed from first ship construction costs, based on ship parameters, cost estimating relationships, material, labor and overhead factors. The fleet construction cost is next computed as a function of the fleet size, the ship production facility characteristics and learning relationships. Fleet costs are a function of ship delivery schedule, and phased by fiscal year. Other contractor investment costs related to the ship and its characteristics include support equipment, spares, training, management, and engineering. Many of these costs will not vary with ship design, and can be expressed as constants for a similar project of the scale of the present one. Many Government investment costs are constants provided by the project office. Operations and support costs are computed on a per ship basis, using operational profile information, ship characteristics and system descriptions (for example, crew size relationships). Next, the fleet costs are computed as a function of ship delivery schedule and fleet size.

Operations and support costs are discounted to properly consider the sacrifice of capital in the civilian sector through commiting of funds to a program over a long term. Many suggestions have been made as to the appropriate value of the discount rate; one very persuasive analysis indicates that it should approximate the average industrial rate of return since this is the product foregone by the civilian sector when operations and support funds are committed to a particular military program. (The foregoing of funds should not be confused with appropriation commitments which are usually made on an

annual basis.) Through discounting of operations and support costs (at 6 percent at FDL, more recently for the LHA Program discounting occurs at 10 percent) an economic measure of a particular system or subsystem configuration results: the sum of development and investment costs together with the discounted present value of operations and support costs. This figure may be compared for alternative systems or subsystem designs in order to select the least life-cycle costs alternative. Figures-of-merit may be computed as well, such as the transportation momentum measure discussed earlier. In Defense planning in the past, it has often been the case that undiscounted operations and support costs are used. In comparing alternatives with unequal lives, this is an inappropriate procedure. In effect, the use of undiscounted costs for a given number of years is the equivalent of using discounted costs for a longer period. For example, the use of 10-year undiscounted operations and support costs is equivalent to the use of 20 years of operations and support cost discounted at 7.75 percent. However, such a "rule-of-thumb" neglects cost stream variation from year to year.

While the selection of a discount rate is made for the purpose of appropriately weighing the economic effects of Defense spending choices as between initial and operating costs, it has strong implications for the outcomes of life-cycle tradeoff analyses. A high discount rate, for example, will significantly reduce the present value of operations and support costs with possible significant design impact. In trading off increased investment in automation against the cost savings through reduced manning, for example, a high discount rate will produce a much smaller investment credit against automation for the saving of one crewman. It has been argued, therefore, that low discount rates should be used. It is the author's view, rather that personnel costs should be carefully evaluated. Many costs need to be more carefully estimated and included in the total military personnel costs. These costs should include not only initial pay and allowances and "fringe benefit" payments, but such costs as the prorated share of equipment used for basic, recruit, and advanced training not particular to a specified weapons system.

IV. SUBSYSTEM TRADEOFFS

At this stage, the overall ship and system parameters have been defined. The ship speed and propulsion plant type has also been specified. The detailed design of the various subsystems of the ship: the hull, propulsion, electric plant, communications and control, auxiliary, outfit and furnishings, and armament must next be elaborated. In order to continue to follow the economic criterion of minimized life-cycle costs subject to side constraints on mission and performance requirements, a subsystem tradeoff procedure (Fig. 8) is used. Not shown in the figure is the way in which candidates for subsystem life-cycle cost tradeoffs are identified nor the way in which design alternatives are selected. Historical data, engineering judgment, and experience are used to analyze the detailed structure of the ship and compare elements of ship structure with elements of life-cycle cost in order to determine those areas where significant life-cycle cost reductions may be effected through the use of the subsystem tradeoff process. With these candidates isolated (a simple rule of thumb might be to define them as subsystems whose cost is a given percentage of total ship construction costs, or whose life-cycle costs are a given percentage of expected total life-cycle cost) a detailed "design work study" procedure is followed to identify in great detail the makeup of these subsystems and major components, the interfaces between them and other subsystems and components of the ship and to identify the critical mission, performance, and engineering factors which have an impact on the selection of a preferred design alternative. Reliability, maintainability, availability, contribution to probability of

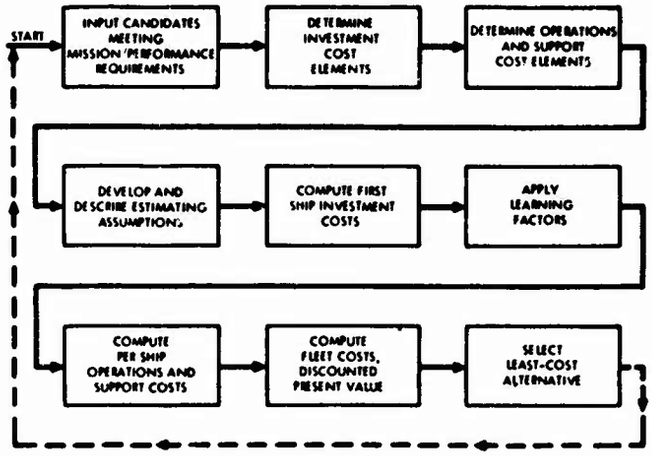


FIGURE 8. Life-cycle cost subsystem tradeoff procedure

mission success, growth potential, safety, and many other factors are considered. Physical performance requirements, such as power and range are considered. Factors such as technical and delivery schedule risk are also assessed. A design work study evaluation matrix is set up with the different design alternatives represented as rows and the different criteria represented as columns. For each criterion, a minimum performance requirement, expressed numerically wherever possible, is specified. Each alternative is then evaluated to see if it meets all requirements. Failure to meet any single requirement is grounds for the redesign or disqualification of that particular alternative. Upon completion of this process, many design alternatives, all meeting mission and performance requirements, are available as input candidates to life-cycle cost tradeoffs. This process helps to separate cost and effectiveness criteria where such a sequential separation is possible. Recall that the basic nature of the cost effectiveness analytic process is such that it is possible to follow one of two pure strategies: a) minimize cost for a fixed effectiveness, or b) maximize effectiveness for a fixed budget. In most government procurements the contractor performs analysis in a competitive environment; it is rare for the government to specify a price and request competition on the basis of maximum effectiveness. Usually the "specified effectiveness-minimize cost" approach is used, allowing competitors to be validated on effectiveness grounds, and evaluated on the basis of their costs; the validation process confirms or refutes the contractors' contention that he has met or exceeded the specified effectiveness requirements. All of his cost predictions are carefully validated following which an evaluation of validated life-cycle costs of alternative offerings makes a selection possible on a least life-cycle cost basis. This is an oversimplification, but it illustrates an important basic principle. In constructing an environment in which contractors are to perform analysis resulting in a system design and specifications, many problems can be avoided through the government determining the effectiveness it requires of a system, and permitting the contractors to then design least life-cycle cost systems meeting this target.

One problem is that of constraining elements which must not be deleted from a system during life-cycle cost analysis. The solution to this problem is to more carefully define, during the concept formulation phase, the values of the various effectiveness measures that the system must meet. Of course, the "rule of reason" applies here. If it is indeed true that one can obtain something for nothing (effectiveness above the minimum required at little or no cost) then contractors should be motivated to seek this effectiveness. This can be done through the appropriate use of weightings in the evaluation criteria for effectiveness above the minimum. These weights should, however, be constructed so that

increases in effectiveness above the minimum requirements, incurred at significant cost, will not be rewarded. Otherwise, the contractor must decide between lower costs or higher effectiveness, usually without any explicit quantitative guidance from the government as to its true wishes. As had been explained earlier, in FDL explicit performance and mission envelopes were specified; these were the equivalent of minimum effectiveness requirements that the system must meet.

The subsystem life-cycle cost tradeoff procedure began with input candidates meeting mission and performance requirements submitted to life-cycle cost analysis. Investment and operations and support cost elements were first determined individually for each tradeoff. The elements of the life-cycle cost structure which would significantly vary between design alternatives were identified and estimating assumptions were developed and described in detail. The parameters for these assumptions were next specified and the appropriate elements of life-cycle cost calculated. Note that the process shown is an iterative process. After selection of the least cost alternative it is possible to develop lower cost elaborations of the least cost alternative, and repeat the tradeoff. In some cases, several alternatives are quite close to each other in total life-cycle costs and the entire tradeoff must be reevaluated, perhaps with careful modification of alternatives. As the ship and system design proceeds in more and more detail many factors, which were assumed, have their values more accurately known. Many details of the ship design become more clearly specified. Thus, it frequently is advisable to repeat tradeoffs although the basic character of the design alternatives may not have changed significantly.

V. SPECIAL STUDIES

Many specialized questions were explored during the FDL Contract Definition through the use of economic analysis and life-cycle cost studies. In some cases, the life-cycle cost tradeoff methodology was applied across many subsystems, as in manning/automation tradeoffs, maintenance and repair resource allocation tradeoffs, and overhaul cycle analyses. The life-cycle cost structure was used to identify all pertinent elements of life-cycle cost and to compare alternatives which had an impact on the balance of cost between these elements. Other studies, such as those related to the production facility design, were conducted using specialized methodology in each case. In the FDL life-cycle cost structure, for example, the facility costs chargeable to the FDL Program made up only one element of the life-cycle cost structure defined by the Navy. One could, beginning from this point, develop a complete life-cycle cost structure for the facility itself. Many detailed and elaborate tradeoffs were conducted to determine the location, configuration, and process flow for the production facility.

One of the many economic analyses that led to the design of our proposed FDL shipyard follows the classic pattern of production function analysis. The simple production function in economics is analogous to the "2 inputs, 1 output" case described frequently in the systems analysis literature. The ideal case (Fig. 9) consists of a series of iso-output curves (isoquants) which describe combinations of capital and labor which would result in a fixed output. For example, the 14-ship isoquant shows those combinations of capital and labor in a shipyard which would result in the capability to produce 14 ships per year. Similar curves for lower output are shown for 12 ships per year and 10 ships per year. Each point on such an iso-output curve represents an efficient combination of capital and labor. That is, for a given capital cost it is assumed that the iso-output curves reflect the least labor cost that, combined with the amount of capital will produce the specified number of ships. The iso-output curves reflect production possibilities. There is no implication that all points on a given iso-output curve reflect a particular total cost, but rather the production of a particular total output.

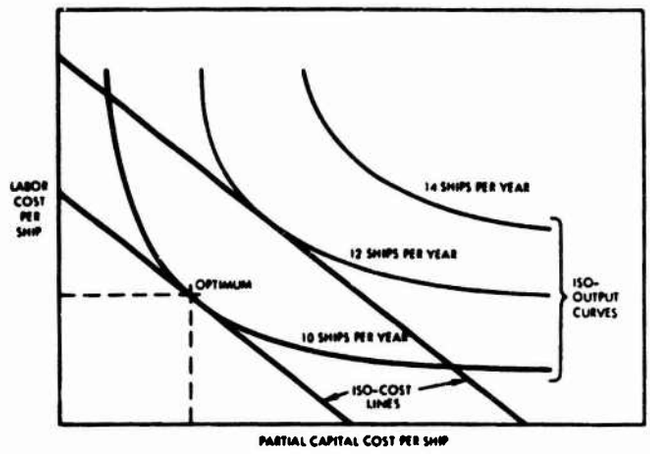


FIGURE 9. Production facility analysis; Ideal case

Isocost lines (budget or exchange curves) are also shown. These represent the amounts of capital and labor that can be purchased for a fixed total cost. We see that the upper isocost line runs from a point on the labor cost axis reflecting the commitment of all financial resources to labor, to a point on the capital cost axis reflecting that same commitment to capital equipment. The isocost line is the locus of all such combinations which have the same total cost. Isocost lines reflect amounts of capital and labor that can be bought for a fixed budget; there is no implication as to the output one can produce at any point on a given isocost line.

If we are interested in producing 10 ships per year (the lowest iso-output curve) then the optimum mix of capital and labor would be that point on the 10-ship iso-output curve which is just tangent to the lowest isocost line. Any smaller total budget will not permit the production of 10 ships per year. A higher isocost line would reflect a larger budget than necessary to produce 10 ships per year. This optimum can be found analytically as well as graphically in many cases, although elaborate computational tools are sometimes required. In the real world, iso-output curves are not so smooth and regular nor are isocost curves necessarily straight lines. This is partly due to the lumpiness of capital: in a major physical facility such as a shipyard, capital is not infinitely divisible and the choice of, for example, ship erection and launch facilities is restricted to a number of discrete possibilities. In an analysis of the optimum ship erection and launch facility for the proposed new shipyard, 120 alternative capital equipment configurations which could produce the required number of ships per year were defined. For each such configuration the labor necessary for efficient use of that capital facility was determined. Labor manhours between alternate flow paths (Fig. 10) varied 9 percent while capital costs varied 40 percent. It is clear that many of the combinations shown are extremely inefficient. In particular, three combinations (the exaggerated dots) clearly resulted in higher labor for a given amount of capital than many of the others in the collection of alternatives. The alternatives were next plotted on appropriately normalized per ship scales with budget curves also shown (Fig. 11). The five alternatives shown were the least labor cost alternatives for the given amounts of capital. It is clear that due to the lumpiness of capital equipment and the inefficiencies of some of the remaining combinations, labor cost did not uniformly decrease as capital cost increased as expected from the theoretical isoquants. In particular, alternatives 1 and 2, while the least labor cost alternatives for the given capital amount, represented "irrational" machinery combinations. Alternatives 12 and 8 were clearly the least cost alternatives in the analysis and were chosen on a basis for further detailed elaborations of the produc-

tion erection and launch scheme, elaborations which were then subjected to more detailed cost tradeoff analysis. The findings of the analysis shown here were quite sensitive to amortization assumptions; the choice between facility design alternatives depends heavily on the amortization that would be permitted over time.

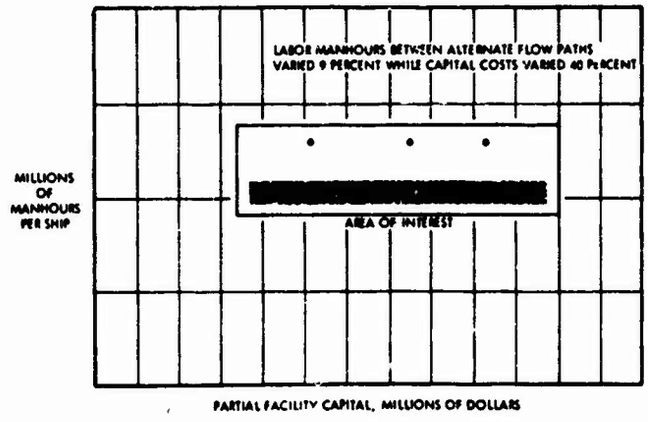


FIGURE 10. Production facility analysis; Ship erection and launch alternatives

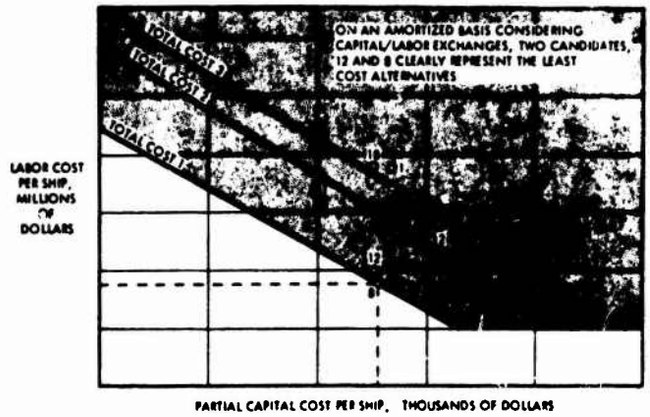


FIGURE 11. Production facility analysis; Ship erection and launch comparison

Extensive studies were conducted of automation and manning. In the individual subsystem tradeoffs, different levels of automation and manning were assumed where appropriate, and suboptimization of subsystem configurations took place through the subsystem tradeoff method. Overall systems optimization, however, considered the fact that both crew members and automation are not infinitely divisible, and different crew and automation functions are complementary goods. In our final manning studies, crew size was determined by considering all the operational, technical and support tasks that the crew of the proposed ship had to perform. Many alternative crews were considered, together with the appropriate level of automation for each crew. For each crew size, the incremental life-cycle cost (both crew and automation-related was determined (Fig. 12). At the time of the analysis, there were uncertainties about regulatory and MSTS requirements for crew size as a function of the ship design. Sensitivity analyses were, therefore, conducted and the upper and lower curves show the band within which the requirements were expected to fall. Automation in the proposed ship, for example, could vary between point 1 and point 2. Automation in current practice is also shown, as are life-cycle cost

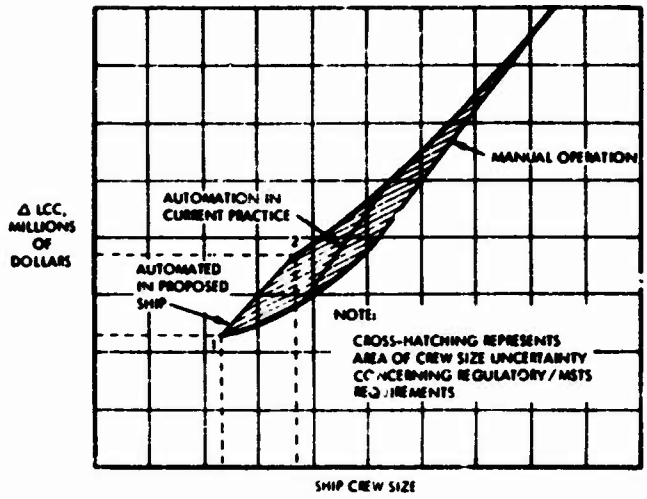


FIGURE 12. Automation vs manning

changes for manual operation of the ship. The exchange between automation and reduced crew size is an extremely attractive one in this range of feasible crew sizes (chosen with due regard to minimum manning and maintenance tasks that must be performed to keep the ship operational). The degree of feasible automation in the proposed ship results in a crew size significantly smaller than that for a ship automated to the level of the best new-design commercial cargo ships.

VI. DETAILED DESIGN

Design below the level of subsystem tradeoffs was conducted by the engineering design groups without the use of formal life-cycle cost tradeoffs. Many hundreds of design decisions are made each day in a project of this kind; it would not be possible to document all of these decisions as formal life-cycle cost tradeoffs when cost was a significant factor. Engineers were given detailed instructional material on life-cycle cost structure, analysis and tradeoffs, and rules of thumb were provided to make it possible to select between alternatives in the absence of complete information. The normal pricing process, selecting between vendors of similar hardware, also permitted cost minimization. Where significant differences did not exist between operations and support costs, selecting the least acquisition cost alternative (the "low bidder") provided for valid decisions. During the pre-production phase of a program of this kind, many of these decisions can be reexamined more carefully in an attempt to achieve still further cost savings. Our experience has revealed that engineers can properly consider significan life-cycle cost factors in making their detailed design decisions. Rules of thumb were developed to aid in these decisions, particularly when an operating cost difference was felt to exist but could not be quantified. The difference in operations and support costs necessary to offset a difference of $1,000 of investment cost was defined. Engineers could frequently determine whether a design alternative having higher investment costs was likely to have operating costs which were comparatively low enough to offset this difference.

VII. SUMMARY

This paper has briefly illustrated the way in which analyses and tradeoffs at many levels in the Contract Definition of a ship and system were used to integrate economic criteria into the process from beginning to end. As a result of our experience with FDL, we have developed methodological and managerial insight into this process, which was used in our successful Contract Definition efforts on the

LHA ship system and the Spruance-class destroyer system. The benefits from life-cycle cost and economic analysis integrated into major physical system planning and design are so significant that we have adapted these same techniques for many other systems which are currently under in-house study and design for both defense and nondefense application. The technique of formally applied, integrated life-cycle cost analysis is being applied by the Defense Department to many current and future procurements including individual items of hardware. From the design of resistors to that of major systems, substantial savings are possible in overall life-cycle costs. At the same time, more reliable, more maintainable systems will be produced, with the higher investment costs fully justified by the reduction in total life-cycle costs. To assure these benefits, contractors must rise to the responsibility of developing data bases on their products' costs and performance. Careful analysis and complete validation of claims for life-cycle cost savings will be required. Finally, with cost and performance incentives and penalties covering the operations and support period of a product's life, time will become the ultimate validator.

STATISTICAL QUALITY CONTROL OF INFORMATION

Irwin F. Goodman

Army Tank-Automotive Command
Warren, Michigan 48090

ABSTRACT

This paper was written to promote interest by management and statistical quality control personnel in the current need for statistical quality control of information of all types. By way of illustration, a step by step procedure for implementing such control on computer files is presented. Emphasis has been placed on the sequencing of the system rather than the underlying techniques.

INTRODUCTION

During the past 50 years a need has been recognized for statistical quality control procedures and techniques in product oriented industries. Another industry product and by-product, "information," is also in need of techniques and procedures of statistical quality control. Many contemporary decisions are dependent upon vast storehouses of information. For parts to fit together, machine and product tolerances must be closely controlled; likewise, to assure valid decisions, the attendant data bases must be subjected to sound statistical quality control.

Decision making processes at the Army Tank-Automotive Command are not unlike other large government and nongovernment industrial enterprises. During the past 15 years a considerable portion of the logistics and engineering effort has been computerized. This resulted in a considerable number of support and reference ADP files that constitute the data input for the computer. The files vary in size from 50,000 records up to millions of records. In terms of alphanumeric characters some of the files have from 50 million to 10 billion characters. The storage of such large quantities of information and the necessary referencing of the files, as often as three to five times a day, has resulted in the necessity for establishing data base validity, purification of the data files, and statistical quality control.

The purpose of this paper is to promote interest of management and quality control personnel in this significant area of statistical quality control of information. Therefore, the following discussion is presented primarily in terms of the necessary steps or tasks involved. The statistical techniques and methods shown here do not give optimum results in terms of sample size requirements and cost benefits. Random sampling, rather than more sophisticated sampling procedures is employed to simplify the presentation. In the following example, a sample size of 900 is obtained. By applying more sophisticated techniques such as stratified sampling, sequential sampling, etc., the 900 required inspections could be reduced considerably.

DATA BASE VALIDITY

In the Statistical Quality Control of Information at the Army Tank-Automotive Command efforts were initially centered around studies to ascertain a measure of the validity of the data in the computer ADP files. These studies involved a comparison of information in the computer file with the source, which was either a hard copy document or another computer file. Inspection criteria were limited to the

following overview data characteristics: match, mismatch, or can't find. These studies provide a yardstick and some directional priority with regard to data base purification. Similar efforts in the literature are reflected in papers by Benz [1], Bryson [2], and Minton [4].

DATA BASE PURIFICATION

The data base purification effort was concerned with an after the fact evaluation of the data in the computer ADP files. This consisted of essentially a technical edit, although it was also concerned with format. Examples of a technical edit are correct stock number, correct nomenclature, correct stratification codes, correct weight data, and correct dates (such as delivery). Format is concerned with such data characteristics as numeric information in a numeric data field, alphabetic information in an alphabetic data field, alpha-numeric information in an alpha-numeric data field, right or left justified entry of information in the data field, and length of the data information entry. Accomplishment of the purification efforts followed by the periodic conduct of validity studies pointed to the need for a quality control effort. This need applied to both the data input and ADP data maintenance, such as the updating of the computer files.

STATISTICAL QUALITY CONTROL PROCEDURE

The purpose of the statistical quality control procedure is to assure that the percent of incorrect data entries in computer data files does not exceed a specified value. The establishment and conduct of a statistical quality control procedure is presented here in terms of portions of a particular computer file. The data and nomenclature have been coded for illustrative purposes. An essential underlying assumption in the procedure is that the "source" information is correct. Therefore, when a particular computer record does not match the source, the computer record is considered in error. There is one exception to this, if there is an entry in the computer record, but no entry in the source, the inspection is considered "can't find".

STEPS IN THE ESTABLISHMENT OF A STATISTICAL QUALITY CONTROL PROCEDURE

The steps necessary for the establishment of a statistical quality control procedure for an ADP computer file are: Description of Data File, Description of Data Source, Inspection Criteria, Sample Size Required, Allocation of Sample, Inspection, and Statistical Computations and Quality Control.

Description of Data File

The initial step is to determine which data elements are to be inspected from the computer records for the computer file that is to be controlled. This requires information regarding the composition of the computer file. Types of data required are data element nomenclature, definition and purpose of the information, identification, location, quantity of characters, and whether the information is alphabetic (A), numeric (N), or alpha-numeric (AN) in the computer file.

For this example, the information in the computer file was maintained on magnetic tape. Printed listings were obtained through a computer interrogation process and used as the document to be inspected.

The data elements to be statistically quality controlled were selected by individuals responsible for the decisions made with the information. Selection was based on the sensitivity of the decisions to the information of the data elements in the computer files. The data elements selected in the current

example are: Contract Number, Federal Stock Number, Item Name, Procurement Request Order Number (PRON), Procurement Request Order Number (PRON) Date, Contract Date, Quantity Shipped, Contract Value, Depot Code, Delivery Date, Accounting Classification Code, Army Management Structure Code, Unit Price, Financial Inventory Accounting Code, Contract Quantity, Supply Status Code, and Procurement Request Order Number (PRON) Quantity.

Description of Data Source

The data source for the current example was determined to be primarily the contract folder with various hard copy documents. They were stored in file cabinets. The file structure is described in Table 1.

TABLE 1. *Contract File Structure*

Geographical partition code	Date	Quantity cabinets	Quantity drawers	Fraction of total
1	1966	17	68	.245
2	1967	13	52	.187
3	1968	1	4	.014
4	1966	7	28	.101
5	1967	14	56	.201
6	1968	1	4	.014
7	1966	4	16	.058
8	1967	5	20	.072
9	1967	½	2	.007
10	1967	¼	1	.004
11	1967	1	4	.014
12	1967	¼	1	.004
13	1967	1½	6	.022
14	1967	1	4	.014
15	1967	1½	6	.022
16	1967	1	4	.014
17	1968	½	2	.007
TOTALS.....		69½	278	1.000

Inspection Criteria

The inspection criteria is divided into two types: Technical and Format. A few examples of format and technical edit criteria are as follows:

Format Criteria:

Data Element	*Criteria*	*Format (F) or Technical (T)*
Federal Supply Class	4 digit Numeric	F
Julian Date	4 digit Numeric	F
Serial Number	Numeric or Alphabetic, but all card columns must be filled	F

Technical Criteria:

Data Element	*Criteria*	*Format (F) or Technical (T)*
Input Code	One of the following: F10, G11, H12, I13, I17, J14, J17, K15, K17, L16, L17, M18, N22	T
Reference Number Action	One of the following: M18, N20, P25, Q26	T

The inspection results were classified as follows:

MATCH: The entry in the computer file record matches the corresponding entry in the source file.

MISMATCH: The entry in the computer file record does not match the corresponding entry in the source file.

OMISSIONS: There is no entry in the computer file record.

CAN'T FIND: There is no entry in the source file.

Sample Size Required

The number of inspections required to determine the percent of data not correct in the computer file depends upon the accuracy requirements for the results as well as the desired confidence associated with this accuracy. Sample size requirements (Ref. [3]) when the accuracy is prescribed in absolute deviations about or in relative percent of a parameter being estimated have been calculated on a computer time sharing terminal using formulae based on the normal approximation to the binomial distribution. A 95 percent confidence level was assumed and the results are presented graphically in Figs. 1 and 2. The results apply when random sampling is employed and can be improved by using more sophisticated techniques as indicated above.

The methodology to determine the sample size required when accuracy is prescribed in absolute deviations, namely,

$$(\pm E \text{ about } P)$$

$$P \pm E$$

$$E = 2\sigma = 2[P(1-P)/N]^{1/2}$$

$$N = 4P(1-P)/E^2.$$

The sample size required when accuracy is prescribed in relative percent, namely,

$$(\pm D\% \text{ of } P)$$

$$P \pm D\%P$$

$$(D/100)P = 2\sigma = 2[P(1-P)/N]^{1/2}$$

$$N = 4(1-P)/(D/100)^2P,$$

where,

N = required sample size,

P = value of parameter being estimated (proportion not correct),

E = prescribed accuracy in absolute deviations (proportions),

D = prescribed accuracy in relative percent, and

2σ = 95 percent confidence limits.

In the current example, assuming the estimated fraction of incorrect data in the computer file is about 0.10 (P) and that it is prescribed that the true value lies somewhere between ± 0.02 of the measured value, then referring to Fig. 1 the required sample size is 900. In this case, the prescribed accuracy, ± 0.02, was stated in absolute deviations, E. The same example can be restated giving the prescribed accuracy in relative percent, D, as follows: Assuming the estimated fraction of incorrect data in the computer file is 0.10 and that the true value lies between ± 20 percent of the measured value, then referring to Fig. 2 the required sample size is 900.

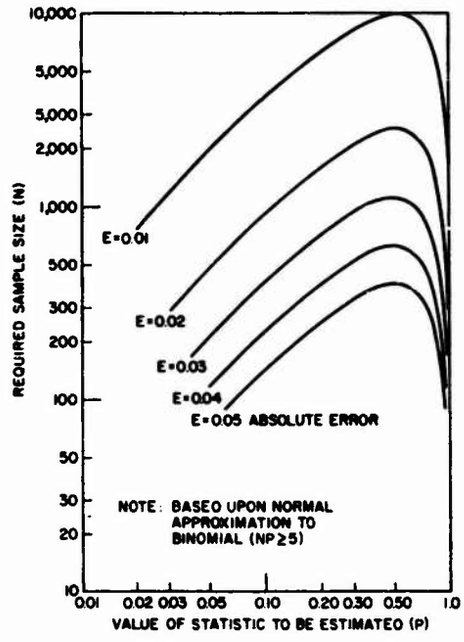


FIGURE 1. Influence on required sample size, N, of required deviation in accuracy (absolute) in parameter, P, being estimated ($P \pm E$ for 95% confidence)

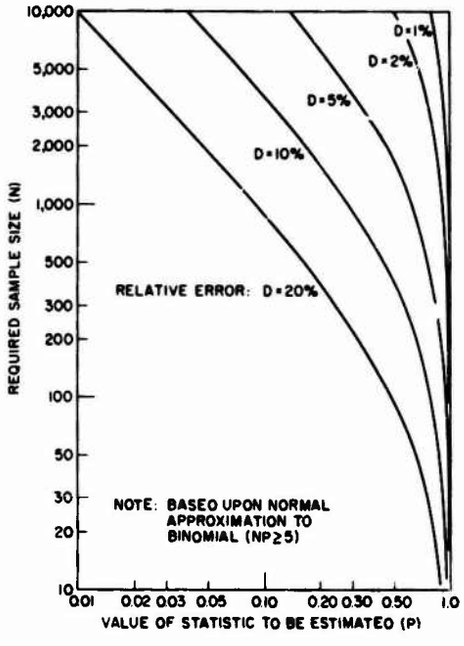


FIGURE 2. Influence on required sample size, N, of required deviation in accuracy (relative) in parameter, P, being estimated ($P \pm D\,\%\,P$ for 95% confidence)

The preceding can be summarized as follows: In order to estimate the fraction incorrect, P, within ± 0.02 in terms of absolute deviations and within ± 20 percent in terms of relative percent, the number of inspections required should be 900.

Allocation of Sample

After the sample size has been established, 900 documents in the current example, then 900 source documents, contract folders, must be randomly selected from all the file cabinets, a considerable undertaking. The file structure was earlier defined to consist of seventeen subgroups classified according to the year and geographic area they represent. The problem of randomly drawing the sample among the subgroups was accomplished by partitioning the sample in proportion to the subgroups. In the example, if 900 is the required sample size and the objective is to randomly sample the 278 file cabinet drawers containing the hard-copy source documents, the allocation of the sample is accomplished as follows: Multiply the "sample size 900" by the "subgroup fraction of total" in the third column of Table 2. The resulting allocation of sample values are shown in the fourth column of Table 2.

TABLE 2. *Allocation of Sample*

Geographical partition code	Quantity of file drawers	Fraction of total	Allocation of sample
1	68	0.245	219
2	52	0.187	168
3	4	0.014	13
4	28	0.101	91
5	56	0.201	180
6	4	0.014	13
7	16	0.058	52
8	20	0.072	65
9	2	0.007	6
10	1	0.004	4
11	4	0.014	13
12	1	0.004	4
13	6	0.022	20
14	4	0.014	13
15	6	0.022	20
16	4	0.014	13
17	2	0.007	6
Total.....	278	1.000	900

After the number of observations to be taken from each of the files has been determined, the particular documents to be selected from the cabinets are determined. This selection process was accomplished with random numbers as follows:

Suppose there are 782 documents in the file, with the partition code 17. Then corresponding to the six observations required for geographical partition code 17, six random numbers were selected in the interval 0 to 782 and the source documents were selected according to their order in the file.

Inspection

For each data element, the inspection consisted of recording and then comparing the data entries in the selected contract folders with the print-outs of the computer ADP files. Work sheets for recording the data entries and making the necessary computations were prepared. The inspection criteria were already discussed above. Briefly summarized there were two types of inspection, format and technical. The results were initially classified as match, mismatch, omissions, and can't find.

Statistical Computations and Quality Control

An example of some inspection results and statistical computations is shown in Table 3. Statistical tests were conducted for significance between results from inspection period to inspection period and also between data elements for a particular file. In addition, the results were usually ranked from high to low in terms of percent not correct.

TABLE 3. *Inspection Results*
(Inspections Attempted for Each Data Element: 900)

Data element	Can't find (b)	Inspections accomplished (c)	Quantity match (d)	Quantity omission (e)	Quantity mismatch (f)	Total not correct (e&f) quantity	Percent not correct (e&f)/c (%)
1	0	900	880	20	0	20	2.2
2	0	900	870	30	0	30	3.3
3	5	895	840	55	0	55	6.0
4	8	892	862	30	0	30	3.4
5	0	900	790	20	90	110	12.0
6	4	896	856	0	40	40	4.5
7	1	899	829	0	70	70	7.7
8	0	900	860	20	20	40	4.4
9	0	900	880	10	10	20	2.2
10	5	895	855	20	20	40	4.5
11	0	900	870	0	30	30	3.3
12	0	900	900	0	0	0	0.0
13	0	900	890	10	0	10	1.1
14	6	894	844	0	50	50	5.5
15	0	900	850	20	30	50	5.5
16	3	897	897	0	0	0	0.0
17	0	900	840	30	30	60	6.6
Total...	32	15,268	14,613	265	390	655	4.3

The results can be further summarized over several sampling periods, as seen in Table 4.

TABLE 4. *Summary of Results*
(In percent)

Result	Period studied							
	1	2	3	4	5	6	7	8
Inspection accomplished [a]	90	95	92	97	94	99	98	99
Match	92.4	91.7	94.9	92.8	90.9	93.6	94.8	95.7
Omission	2.1	2.5	1.9	2.4	2.2	2.4	1.9	1.7
Mismatch	5.5	5.8	3.2	4.8	6.9	4.0	3.3	2.6
Not correct	7.6	8.3	5.1	7.2	9.1	6.4	5.2	4.3

[a] Attempted less can't find.

The results of the periodic inspection are then graphed in quality control chart format. Such charts were prepared for selected data elements, as well as for all the data elements studied. Using the above data, an example of a typical quality control chart is shown in Fig. 3.

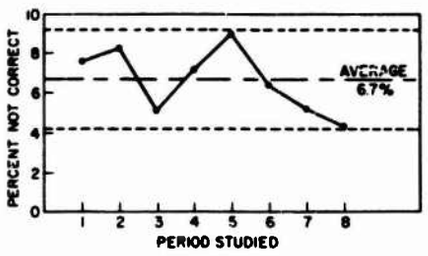


FIGURE 3. Statistical quality control chart (3σ confidence limits)

FUTURE DIRECTIONS

The future directions of Statistical Quality Control of Information should include computerizing the inspecting process (Ref. [5]) the statistical computations, and automatically portraying a statistical quality control picture of the results. Another direction for research could involve the establishment of a decision making matrix showing the data elements necessary for each of the decisions and dynamic indicators reflecting the goodness potential of the decisions due to changes in validity in the data base. Improved sampling and allocation procedures would also be very beneficial.

CONCLUSIONS

In conclusion, it is hoped this paper will promote interest of management and quality control personnel in this new and much needed area of statistical quality control of information. Currently only a dearth of literature exists relevant to the subject.

REFERENCES

[1] Benz, William M., "Quality Control in the Office," *Industrial Quality Control* **23**, 531–535 (May 1967).

[2] Bryson, Marion R., "Practical Application of Operations Research in Physical Inventory Control," *1961 SAE International Congress and Exposition of Automotive Engineering* (Cobo Hall, Detroit, Michigan, 1961).

[3] Cochran, William G., *Sampling Techniques* (John Wiley & Sons, New York, 1965).

[4] Minton, George, "Inspection and Correction Error in Data Processing," *Am. Statist. Assoc. Jour.* **64**, 1256–1275 (Dec. 1969).

[5] O'Reagon, Robert T., "Relative Costs of Computerized Error Inspection Plans," *Am. Statist. Assoc. Jour.* **64**, 1245–1255 (Dec. 1969).

STOCHASTIC DUELS WITH LETHAL DOSE

N. Bhashyam

Defence Science Laboratory,
Delhi-6, India

ABSTRACT

This paper introduces the idea of lethal dose to achieve a kill and examines its effect on the course and final outcome of a duel. Results have been illustrated for a particular case of exponential firing rates.

INTRODUCTION

Williams and Ancker [3] developed a new model to study combat situation by considering it as a two person duel and incorporating in the analysis the microscopic aspects of a combat. The model has since been termed the *Theory of Stochastic Duels.* The details of work done by various analysts in this topic are contained in Ancker [1].

In the various studies conducted so far it has been assumed that a single success by the duelist ensures his win. This assumption, as we shall see presently is valid only in the following cases:

(a) The target, which happens to be the opposing duelist, is such that one hit alone is sufficient to destroy it.

(b) The quantity of ammunition delivered per round is at least equal to or more than the lethal dose required to completely annihilate the opponent. This could be the case with heavy guns etc.

The present paper attempts to study a duel situation wherein the opponent cannot be killed by a single successful shot. On the other hand, the kill requires a finite number of hits. This assumption stems from the nature of modern combat. Present day combat is characterized by emphasis on heavy protective armor and cover designed to provide protection and safety to the combatant so that he can effectively continue in the duel. Under such circumstances it is imperative that the quantity of ammunition delivered on the opponent should be sufficient not only to kill the opponent, but at the same time it must also be able to nullify the affects of protection.

A similar situation arises in an air battle. It may not be very appropriate to assume that a single hit alone will be able to bring down the opposing aircraft unless the hit has been at a very critical part of the aircraft like the fuselage. In order to be able to bring down the aircraft, it will be plausible to assume that we succeed in repeatedly hitting it, which will ultimately force it to go down.

STATEMENT OF THE MODEL

These considerations have been incorporated in the present paper by introducing the idea of lethal dose. We assume that two contestants A and B, each with an unlimited supply of ammunition, are locked in a duel.

Let X_n be a continuous positive random variable denoting the elapsed time since duelist A has fired his nth round. Then $\{X_n\}$ is a sequence of identically distributed independent positive random variables with a density function $D(x)$, such that

$$Pr(X_n \leq x) = \int_0^x D(x)dx.$$

Further, let $\lambda(x)dx$ be the first order conditional probability that A will fire a round in the interval $(x, x+dx)$ given that he has not fired prior to time x. Obviously

$$D(x) = \lambda(x) \exp\left(-\int_0^x \lambda(x)dx\right)$$

Each round fired by A has a probability p of hitting the opponent B and with probability q, A misses B, so that $p+q=1$. Further, it is assumed that each round fired by A delivers a certain amount of ammunition and to kill B a certain fixed quantity of ammunition is required to be delivered by A on B. Let this quantity of ammunition, the *lethal dose*, be contained in R rounds. A kill is said to have been achieved by A as soon as A scores R hits on B.

Similar assumptions hold for duelist B, whose parameters are represented by placing an asterisk (*) as a superscript.

FORMULATION AND SOLUTION

Let us define the following discrete random variables:

$N(t)$: Number of rounds fired by A prior to time t

$$N(t) \geq 0$$

$\theta(t)$: Number of hits secured by A on B prior to time t

$$0 \leq \theta(t) \leq R$$

We now define the following state probabilities

$$P_n^r(x, t)dx = Pr[N(t)=n, \theta(t)=r, x < X_n \leq x+dx | N(0)=\theta(0)=0]$$

$$A_n(t) = Pr[N(t)=n, \theta(t)=R | N(0)=\theta(0)=0].$$

Obviously,

$$P_n^r(x, t)dx = 0 \qquad for\ r > n$$

and

$$A_n(t) = 0 \qquad \text{for } n < R.$$

By continuity arguments we set up the following system of difference-differential equations:

(1) $$\left[\frac{\partial}{\partial x} + \frac{\partial}{\partial t} + \lambda(x)\right] P_n^r(x, t) = 0, \qquad 0 \leq r \leq R-1, n \geq r$$

(2) $$P_n^r(0, t) = q \int_0^\infty P_{n-1}^r(x, t)\lambda(x)dx + p \int_0^\infty P_{n-1}^{r-1}(x, t)\lambda(x)dx, \qquad 1 \leq r \leq R-1, n \geq r$$

(3) $$P_n^0(0, t) = q \int_0^\infty P_{n-1}(x, t)\lambda(x)dx, \qquad n > 1$$

(4) $$E_n(t) = \frac{d}{dt} An(t) = p \int_0^\infty P_{n-1}^{R-1}(x, t)\lambda(x)dx, \qquad n \geq R.$$

Initially

$$P_n^r(x, 0) = \delta_{r,0}\delta_{n,0}\delta(x) \tag{5}$$

where $\delta_{i,j}$ is Kronecker's delta and $\delta(x)$ is Dirac delta function.

We define the following generating functions

$$F(x, t, \alpha, \beta) = \sum_{r=0}^{K-1} \beta^r \sum_{n=0}^{\infty} \alpha^n P_n^r(x, t)$$

$$k(t, \alpha) = \sum_{n=0}^{\infty} \alpha^n E_n(t).$$

Applying the above generating functions to equations (1) to (5), we get

$$\left[\frac{\partial}{\partial x} + \frac{\partial}{\partial t} + \lambda(x)\right] F(x, t, \alpha, \beta) = 0, \tag{6}$$

$$F(0, t, \alpha, \beta) = \alpha q \int_0^{\infty} F(x, t, \alpha, \beta)\lambda(x)dx + \alpha\beta p \int_0^{\infty} F(x, t, \alpha, \beta)\lambda(x)dx - \beta^K k(t, \alpha), \tag{7}$$

and

$$F(x, 0, \alpha, \beta) = \delta(x). \tag{8}$$

Taking Laplace transform and denoting the Laplace transform of probabilities by placing a bar as superscript i.e. $\bar{F}(\mathcal{s}) = \int_0^{\infty} \exp(-\mathcal{s}t)F(t)dt$, $Re\,\mathcal{s} \geq 0$, equations (6) to (8) give

$$\left[\frac{\partial}{\partial x} + \mathcal{s} + \lambda(x)\right] \bar{F}(x, \mathcal{s}, \alpha, \beta) = \delta(x) \tag{9}$$

$$\bar{F}(0, \mathcal{s}, \alpha, \beta) = \alpha q \int_0^{\infty} \bar{F}(x, \mathcal{s}, \alpha, \beta)\lambda(x)dx + \alpha\beta p \int_0^{\infty} \bar{F}(x, \mathcal{s}, \alpha, \beta)\lambda(x)dx - \beta^K \bar{K}(\mathcal{s}, \alpha). \tag{10}$$

Solving equation (9) we get,

$$\bar{F}(x, \mathcal{s}, \alpha, \beta) = [1 + \bar{F}(0, \mathcal{s}, \alpha, \beta)] \exp\left(-\mathcal{s}x - \int_0^x \lambda(x)dx\right).$$

Substituting the value of $\bar{F}(x, \mathcal{s}, \alpha, \beta)$ in (10) we get

$$1 + \bar{F}(0, \mathcal{s}, \alpha, \beta) = \frac{1 - \beta^K \bar{K}(\mathcal{s}, \alpha)}{1 - \alpha(q + \beta p)\bar{D}(\mathcal{s})}. \tag{11}$$

The left hand side of (11) is regular on and inside of $|\beta| \leq 1$, for $Re\,\mathcal{s} \geq 0$ and $|\alpha| \leq 1$. In this domain the denominator of the right hand side has a simple zero at $\beta = \hat{\beta}$ where

$$\hat{\beta} = \frac{1}{\alpha p \bar{D}(\mathcal{s})}[1 - \alpha q \bar{D}(\mathcal{s})].$$

Therefore, $\beta = \hat{\beta}$ must also be a root of the numerator, so that

$$\bar{K}(\mathcal{s}, \alpha) = \left[\frac{\alpha p \bar{D}(\mathcal{s})}{1 - \alpha q \bar{D}(\mathcal{s})}\right]^K. \tag{12}$$

Whence, $\bar{H}(s)$, the Laplace transform of $H(t)$, the probability for the time taken by A to kill B, is

$$\bar{H}(s) = [\bar{K}(s, \alpha)]_{\alpha=1}$$

$$= \left[\frac{p\bar{D}(s)}{1-q\bar{D}(s)}\right]^R. \tag{13}$$

Similarly $\bar{G}(s)$, the Laplace transform of $G(t)$, the probability density for the time taken by duelist B to kill A, is obtained as

$$\bar{G}(s) = \left[\frac{p^*\bar{D}^*(s)}{1-q^*\bar{D}^*(s)}\right]^{R^*}.$$

EVALUATION OF WIN PROBABILITIES

Let P(A) be the probability that A wins the duel; then

$$P(\mathrm{A}) = \int_{t=0}^{\infty} H(t) \int_{\tau=t}^{\infty} G(\tau) d\tau dt \tag{15}$$

We know

$$H(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \bar{H}(s) \exp(st) ds \tag{16}$$

Where the path of integration is parallel to the imaginary axis, c being chosen so that all the singularities of $\bar{H}(s)$ lie to the left of the line of integration and $\bar{H}(s)$ is analytic to the right of it.

From (15) and (16) we have

$$P(A) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \bar{H}(s) \left[\int_{t=0}^{\infty} \exp(st) \int_{\tau=t}^{\infty} G(\tau) d\tau dt\right] ds$$

$$= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \bar{H}(s) \bar{G}(-s) \frac{ds}{s} - \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \bar{H}(s) \frac{ds}{s} \tag{17}$$

To evaluate the integral in (17), we choose a semi-circular contour wholly lying on the right of the imaginary axis in the complex plane as shown in Fig. 1

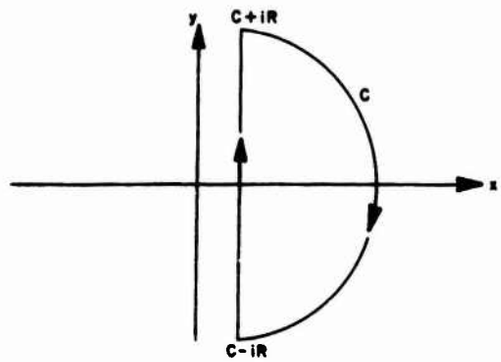


FIGURE 1. Evaluation of integral.

The line is such that it separates the poles of $\bar{H}(s)$ from those of $\bar{G}(-s)$. The poles of the integrand lying in the chosen contour are those belonging to $\bar{G}(-s)$. Hence the second integral in (17) is zero as $\frac{1}{s}\bar{H}(s)$ is analytic everywhere to the right of the line $c-iR$ to $c+iR$. Thus

$$P(\mathrm{A}) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \bar{H}(s)\bar{G}(-s)\,\frac{ds}{s}. \tag{18}$$

It may be remarked here that as $H(t)$ and $G(t)$ are probability density functions, the integral of $\frac{1}{s}\bar{H}(s)$ and $\frac{1}{s}\bar{H}(s)\bar{G}(-s)$ on C (Fig. 1) tend to zero as $R\to\infty$. Thus

$$P(\mathrm{A}) = -\sum_i R_i \tag{19}$$

where R_i is the residue at the ith pole of the integrand in (18) and summation is over all the poles lying inside the contour.

Similarly $P(\mathrm{B})$, the probability that B wins the duel is given by

$$P(\mathrm{B}) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \bar{G}(s)\bar{H}(-s)\,\frac{ds}{s} \tag{20}$$

$$= -\sum_j R_j^* \tag{21}$$

where R_j^* is the residue at the jth pole of the integrand in (20) and summation is over all the poles lying in the contour as in Fig. 1.

THE CASE WHEN R AND R^* ARE RANDOM VARIABLES

Let us now consider the case when the exact number of rounds required to secure a kill is not fixed, but there is a probability distribution giving the number of rounds required to kill. Let

$$Pr(R=m)=\alpha_m$$

such that

$$\sum_{m=0}^{\infty}\alpha_m = 1, \qquad \alpha_0 = 0.$$

Similarly,

$$Pr(R^*=k)=\beta_k$$

where

$$\sum_{k=0}^{\infty}\beta_k = 1, \qquad \beta_0 = 0.$$

Then $\bar{H}(s)$ and $\bar{G}(s)$, the Laplace transforms of the probability densities for the times taken to kill by A and B respectively are given by

$$\bar{H}(s) = \sum_{m=1}^{\infty}\alpha_m\left[\frac{p\bar{D}(s)}{1-q\bar{D}(s)}\right]^m \tag{22}$$

and

$$\bar{G}(s) = \sum_{k=1}^{\infty}\beta_k\left[\frac{p^*\bar{D}^*(s)}{1-q^*\bar{D}^*(s)}\right]^k. \tag{23}$$

PARTICULAR CASES

CASE 1 Inter-firing times exponentially distributed for both duelists:

Let

$$\bar{D}(s)=\frac{\lambda}{\lambda+s},$$

$$\bar{D}^*(s)=\frac{\lambda^*}{\lambda^*+s}.$$

From (13) and (14) we get

$$\bar{H}(s)=\left(\frac{\lambda p}{\lambda p+s}\right)^R$$

and

$$\bar{G}(s)=\left(\frac{\lambda^* p^*}{\lambda^* p^*+s}\right)^{R^*}.$$

Substituting the value of $\bar{H}(s)$ and $\bar{G}(s)$ in (18) we get

$$P(\mathrm{A})=\frac{(\lambda p)^R(\lambda^* p^*)^{R^*}}{2\pi i}\int_{c-i\infty}^{c+i\infty}\frac{ds}{s\,(\lambda p+s)^R(\lambda^* p^*-s)^{R^*}}.$$

Integrating around the contour as in Fig. 1 we find that the integrand has a pole of order R^* at $s=\lambda^* p^*$. We evaluate the residue by collecting the co-efficient of $(s-\lambda^* p^*)^{-1}$ in the expansion of the integrand and finally we obtain

$$P(\mathrm{A})=\left[\frac{\lambda p}{\lambda p+\lambda^* p^*}\right]^R\sum_{j=0}^{R^*-1}\binom{R+j-1}{j}\left[\frac{\lambda^* p^*}{\lambda p+\lambda^* p^*}\right]^j$$

$$=1-I_{\frac{\lambda^* p^*}{\lambda p+\lambda^* p^*}}(R^*, R)$$

where $I_x(p, q)$ is the well tabled (Pearson [2]) Incomplete Beta-Function Ratio defined by

$$I_x(p, q)=\frac{B_x(p, q)}{B(p, q)}$$

$$=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^x y^{p-1}(1-y)^{q-1}dy.$$

Using the relationship $I_x(p, q)=1-I_{1-x}(q, p)$ we get

$$P(\mathrm{A})=I_{\frac{\lambda p}{\lambda p+\lambda^* p^*}}(R, R^*) \tag{24}$$

Similarly,

$$P(\mathrm{B})=I_{\frac{\lambda^* p^*}{\lambda p+\lambda^* p^*}}(R^*, R) \tag{25}$$

Putting $\mathrm{K}=\frac{\lambda^* p^*}{\lambda p}$ in (24) we get

$$P(\mathrm{A}) = I_{\frac{1}{1+K}}(R, R^*)$$

The product λp gives the rate at which A hits B. Similarly $\lambda^* p^*$ is the rate at which B hits A, so that K is the ratio of the hitting rates of the opposing duelists. Graphs have been drawn in Fig. 2 to show the influence of R and R^* on $P(\mathrm{A})$ for $K = \frac{1}{2}, 1, 2$.

CASE 2 Exponential inter-firing times and geometric R and R^*:

Let

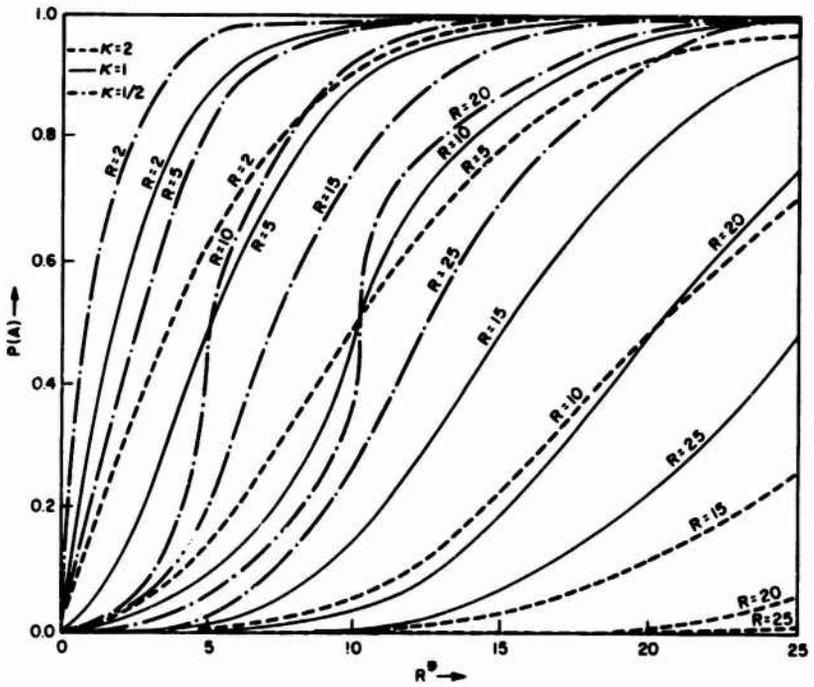


FIGURE 2. Effect of lethal dose on win probability.

$$\bar{D}(s) = \frac{\lambda}{\lambda + s}; \bar{D}^*(s) = \frac{\lambda^*}{\lambda^* + s},$$

$$\alpha_m = (1-\alpha)\alpha^{m-1}$$

and

$$\beta_K = (1-\beta)\beta^{K-1}$$

From (22) and (23)

$$\bar{H}(s) = \frac{(1-\alpha)\lambda p}{\lambda p(1-\alpha) + s}$$

and

$$\bar{G}(s) = \frac{(1-\beta)\lambda^* p^*}{\lambda^* p^*(1-\beta) + s}$$

so that from (18)

$$P(\mathrm{A}) = \frac{(1-\alpha)(1-\beta)\lambda p \lambda^* p^*}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{ds}{s[\lambda p(1-\alpha) + s][\lambda^* p^*(1-\beta) - s]}.$$

Integrating around a contour as in Fig. 1 we find the integrand has a simple pole at $\omega = \lambda^* p^*(1-\beta)$. Hence

$$P(\mathrm{A}) = \frac{\lambda p(1-\alpha)}{\lambda p(1-\alpha) + \lambda^* p^*(1-\beta)}. \tag{26}$$

Similarly,

$$P(\mathrm{B}) = \frac{\lambda^* p^*(1-\beta)}{\lambda p(1-\alpha) + \lambda^* p^*(1-\beta)}. \tag{27}$$

Putting $K = \dfrac{\lambda^* p^*}{\lambda p}$ in (26) we get

$$P(\mathrm{A}) = \frac{(1-\alpha)}{(1-\alpha) + k(1-\beta)}.$$

In Fig. 3 graphs have been drawn to show the effect of α on $P(\mathrm{A})$ for different values of β and K.

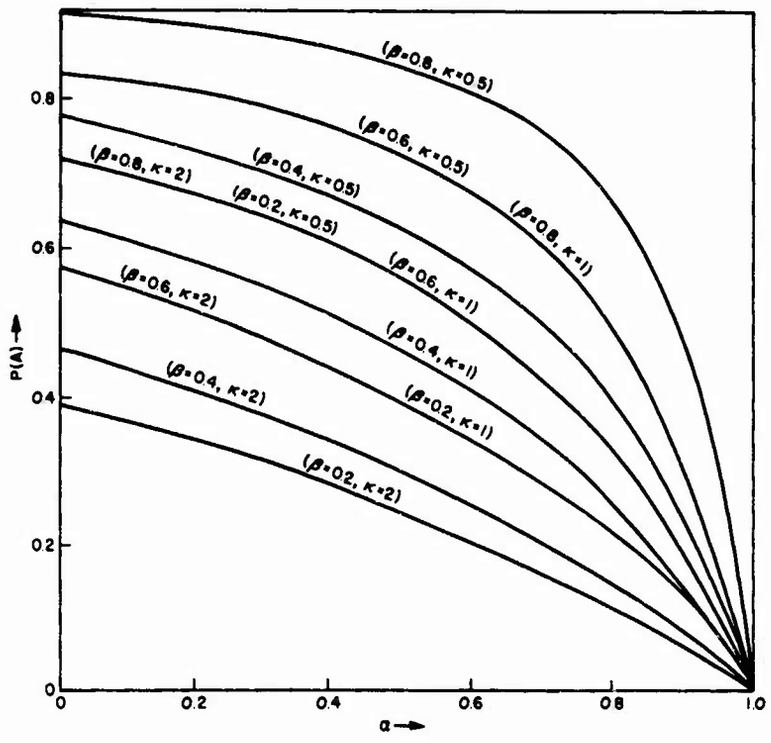


FIGURE 3. Effect of random lethal dose on win probability.

ACKNOWLEDGEMENTS

Thanks are due to Dr. Kartar Singh, Director, Defence Science Laboratory, Delhi for permission to publish this paper. Author is extremely grateful to Prof. R. S. Varma, Dean, Faculty of Mathematics, Delhi University and Dr. N. K. Jaiswal, Head, Statistics and Operational Research Division, Defence Science Laboratory, Delhi for actively guiding him throughout the preparation of this paper.

REFERENCES

[1] Ancker, C. J., Jr., "The Status and Development in the Theory of Stochastic Duels," Opns. Res. **15**, 388–406 (1967).

[2] Pearson, K., "Tables of the Incomplete Beta Function," University Press, Cambridge, 1956.

[3] Williams, Trevor and C. J. Ancker, Jr., "Stochastic Duels," Opns. Res. **11**, 803–817 (1963).

A NOTE ON A PROBLEM OF SMIRNOV
A GRAPH THEORETIC INTERPRETATION

Ronald Alter

University of Kentucky

and

Bennet Liemz

System Development Corporation

ABSTRACT

This paper considers a graph theoretic interpretation of a problem proposed by Smirnov.

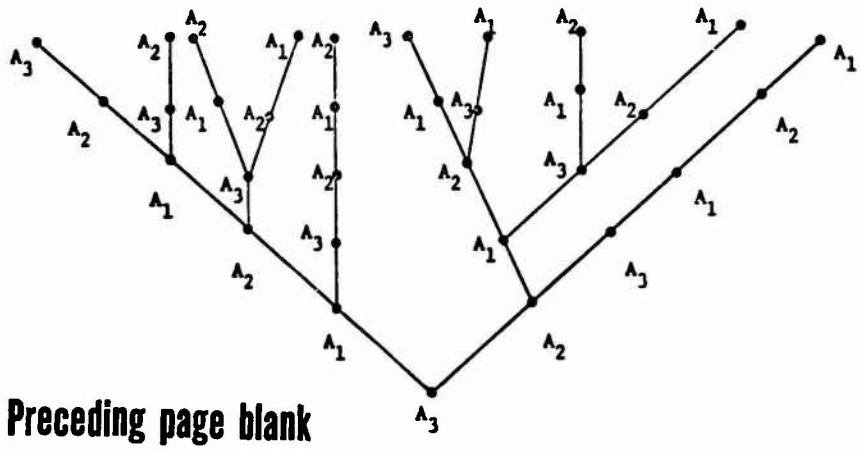
1. INTRODUCTION

The basic problem stated by Smirnov is the following: how many ways can n objects of $s + 1$ classes be arranged in a chain so that no two objects of the same class are adjacent? In Ref. [4] Sarmanov and Zaharov viewed the problem as one of transitions between classes. They obtained limiting results for the case of $s = 2$ (i.e., three classes of objects) and for the case wherein all classes have the same number of objects. These results are summarized in Ref. [2]. The purpose of this paper is to interpret the problem in terms of graph theory and the theory of trees.

2. A GRAPH THEORIC INTERPRETATION

Suppose there are n objects divided into $s + 1$ distinct classes with r_I as the number of objects in the Ith class. Within a given class, all objects are assumed to be indistinguishable. Let $M^{(s+1)}(r_1, \ldots, r_{s+1})$ denote the number of arrangements or chains possible such that no two objects of the same class are adjacent.

It is assumed that the reader is familiar with the usual definitions of graph, connected graph, cyclic graph, and acyclic graph. These definitions appear in Ref. [3]. Using the standard graph theory terminology, a tree is a connected acyclic graph. If a special vertex has been selected as the beginning of the tree T, then this vertex is said to be the root of T, and T is called a rooted tree.



For the purposes of this paper, the drawing of a tree provides a very useful tool for the analysis of the various logical probabilities which arise. The following example serves to illustrate this interpretation.

Example. Let $n = 6$, $s = 2$, and $r_1 = r_2 = r_3 = 2$. Let an object from the Ith class be labeled A_I. Because of symmetry it suffices to consider the root of the tree beginning with an A_3 say and then multiply the total number of chains by 3. One has that which gives $M^{(3)}(2, 2, 2) = 3 \cdot 10 = 30$.

Note: If, for example, $n = 9$, $s = 2$ and $r_1 = 2$, $r_2 = 3$, $r_3 = 4$, then three trees would be constructed, and $M^{(3)}(2, 3, 4) = 79 =$ the sum of the terminal vertices of all three trees.

Several results that are applicable to the theory of trees can now be given, along with some relevant definitions.

Definition 1: A uniform n-tree is a tree in which the shortest path from the root to each terminal vertex is n.

Definition 2: A chromatic tree for colored graphs is a tree in which no two adjacent vertices have the same color.

Thus, interpreting the combinatorial problem graph theoretically it is evident that the problem lies in chromatic uniform n-trees.

Suppose one draws chromatic uniform n-trees in the way described in Examples 1 and 2. Given are n and $s+1$ distinct classes with the Ith class containing r_I objects $n = \sum_{I=1}^{s+1} r_I$. By selecting a representative from each class to a root of one tree, it can be seen that there are $s+1$ trees and that

$$M^{(s+1)}(r_1, \ldots, r_{s+1}) = \sum_{I=1}^{s+1} B_I,$$

where B_I is the number of terminal vertices on the tree whose root is chosen from the Ith class. (Note: In the notation of Ref. [1]

$$B_I = M^{(s+1)}(r_1, \ldots, r_{s+1}; I).).$$

The quantities can be obtained by the methods given in Refs. [1] and [2].

REFERENCES

[1] Alter, Ronald and B. P. Lientz, "A Generalization of a Combinatorial Problem of Smirnov," System Development Corporation, SP-3254.

[2] Alter, Ronald and B. P. Lientz, "Applications of a Generalized Combinatorial Problem of Smirnov," Nav. Res. Log. Quart. **16**, 543–547 (1969).

[3] Harary, F., *Graph Theory* (Addison Wesley Co., Reading, Pa., 1969).

[4] Sarmanov, O. V. and V. K. Zaharov, "A Combinatorial Problem of Smirnov," Dokl. Akad. SSSR **176**, 1147–1150 (1967).

NEWS AND MEMORANDA

NATO OPTIMIZATION CONFERENCE, JULY 1971

A Conference on Applications of Optimization Methods for Large-Scale Resource-Allocation Problems will be held in Elsinore, Denmark, July 5–9, 1971. The Conference is sponsored by the NATO Science Committee and is under the Scientific Directorship of Professors George B. Dantzig and Richard W. Cottle, Stanford University. Attendance will be limited to 120 persons.

The purpose of the Conference is to review and to advance the art of optimizing large-scale resource-allocation problems. Topics of interest include methodology for solving structured mathematical programs, models for national planning, experience with solving large-scale systems, and the need for experimentation.

Readers of this notice are urged to express their interest in participating or in contributing a paper (30 minutes). Abstracts of contributed papers must be received no later than January 30, 1971. Abstracts should be addressed to Professor Richard W. Cottle, Department of Operations Research, Stanford University, Stanford, California 94305.

Dr. Murray A. Geisler, The RAND Corporation, 1700 Main Street, Santa Monica, California 90406, is the American point of contact. Inquiries regarding the Conference may be addressed to him.

U.S. GOVERNMENT PRINTING OFFICE : 1970 – OL-404-971

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH LOGISTICS QUARTERLY

SEPTEMBER 1970
VOL. 17, NO. 3
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217