IR No. 70-36

INFORMAL REPORT

CRITICAL STUDY OF APPROXIMATING FUNCTIONS (AND METHODS) AS APPLIED TO OCEAN STATION DATA PROJECT I REGRESSION ANALYSIS

MAY 1970

NAVAL OCEANOGRAPHIC OFFICE
WASHINGTON, D. C. 20390

INFORMAL REPORT

The Informal Report (IR) as produced at the Naval Oceanographic Office is a means for personnel to issue timely scientific and technical preliminary reports of their investigations. These are primarily informal documents used to report preliminary findings or useful byproducts of investigations and work to members of the scientific and industrial communities.

Informal Reports are assigned sequential numbers for each calendar year; the digits preceding the dash indicate the year.

The distribution made of this report is determined primarily by the author. Information concerning obtaining additional copies or being placed on a distribution list for all future Informal Reports in a given area of interest or specialty field, should be obtained from:

Field Management and
Dissemination Department
Code 4420
Naval Oceanographic Office
Washington, D.C. 20390

ABSTRACT

The stepwise multiple regression technique is used in a model building process to develop predictors of temperature, salinity, and sound velocity as functions of geographical location, time, and depth. Models which give reasonable results are obtained through successive trials using higher order terms of the independent variables. The model for sound velocity yields values which are nearly identical to the Wilson sound velocities contained in the ocean station file and values computed using a modified version of the MacKenzie equation.

The distribution of residuals resulting from comparisons of the Wilson equation sound velocities to those obtained from the regression model (both computed from actual temperature and salinities) shows that 98% fall within the range of ±2 m/sec. A comparison of the regression model sound velocity values computed from regression predictions of temperature and salinity with the Wilson values shows that 88% of the residuals fall in the range of ±12 m/sec.

The results, which are valid for the 4° square centered at 37.5° North latitude and 69.5° West longitude, are discussed in terms of the statistical significance of the distribution of the residuals. Since the physical characteristics of the area selected are rather complex, the application of this technique to other parts of the ocean is recommended.

This work was performed under NAVOCEANO Contract No. N62306-68-C00241 by Dr. Billy E. Gillett, Department of Statistics and Applied Mathematics, University of Missouri in Rolla, Missouri.

APPROVED FOR RELEASE: K. L. Wenckus

Acting Director
Exploratory Oceanography Division

DATE: 1 May 1970

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

In the field of oceanography, there is a real need for accurate numerical procedures for determining sound velocity in sea water based on certain ocean variables such as latitude, longitude, temperature, salinity, and day-of-year.

Although the technique of stepwise multiple regression readily lends itself as a tool of analysis in the development of polynomial prediction equations of the form

$$y = \sum_{i=o}^{n} \beta_i x_i \qquad \text{where } x_i = f(z_1, z_2, \ldots, z_n)$$

where the β_i are the coefficients to be determined and the z_i are the independent variables in the model, no reliable numerical method exists which eliminates the need for on-location measurements of certain variables such as temperature, salinity and pressure. Once the values of these variables are known, however, one may use one of a number of well known reliable equations for computing sound velocity. Two such equations utilized in this study are those of H. V. Mackenzie[1] and Wayne D. Wilson.[2]

It is the purpose of this study to adequately predict sound velocity at any location within a given range of latitude and longitude without going to that particular location to measure variables such as temperature and salinity. In order to do this, however, it is required that temperature and salinity be predicted to a certain degree of accuracy. This will involve examining a number of classes

of models.

The problem of developing prediction equations for temperature, salinity, and sound velocity is further complicated by other factors, most of which are uncontrollable. Some of these factors are time series autocorrelation in the data, errors due to instrumentation, missing data, land masses, underwater streams or currents, temperature inversions, and sparse data, to mention a few. All of these factors have their individual effects on the generalized regression development. The effects of some of these factors will be discussed in the following chapter. It is hoped, of course, that errors due to these factors will occur randomly.

When dealing with oceanographic problems, the handling of data becomes an obstacle. While the data for a given square (x° by x°) is relatively sparse, the total amount of data for this square is extremely voluminous. Consequently, most of the conclusions of this study are based on data from the 4° by 4° square 36° - 40°N latitude and 68° - 72° west longitude. The convention used is as follows: North latitude is positive; west longitude is negative.

The execution time for the stepwise multiple regression procedure when a large model is under consideration is quite long. For purposes of economy the goal is to determine a model with as few terms as possible which does an adequate job of predicting. This requires extensive trial and error model refinement.

Stringent accuracy requirements are needed to qualify the regression analysis as an acceptable subsystem to the more extensive, overall "Ocean Station Display System" and the "quick look" facility utilizing a cathode ray tube, now under development by Mr. Richard Bolton. The regression equations, which are surfaces when plotted, can be instrumental in the display of temperature, salinity, and sound velocity contours in the graphic display system.

Initially, a simple model will be considered at each of the depth planes in the 4° by 4° square 36° - 40° N latitude and 68° - 72°W longitude. This will yield a set of regression equations for temperature, salinity, and sound velocity for each depth plane consisting of terms not rejected by the predetermined accuracy criterion.

A more general regression situation is then considered where an equation is developed using depth as one of the independent variables. This results in the development of one equation for each of the dependent variables temperature, salinity and sound velocity, which is general for all depth planes.

Many general regression models involving as many as six independent variables with up to tenth order cross products were tried. The process of developing the models involved trial and error addition and deletion of cross products of the various independent variables. Several interesting combinations were tried and the models which produced the best results are discussed in the latter part of Chapter III.

The primary objective of this study is to develop equations which may be used to predict temperature and salinity using only controllable variables which may be set by the user. Once these values are known, they may be used in some sound velocity equation such as Wilson's, Mackenzie's, or the regression sound velocity equation.

Once the temperature and salinity equations are developed, five sound velocity calculations are possible for each observation card. Given latitude, longitude, and depth, a temperature and salinity may be calculated from the respective regression equations. This allows calculation of sound velocity from Mackenzie's equation and the regression sound velocity equation using the predicted temperature and salinity. Two more sound velocities may be obtained at this observation by evaluating these two equations using the observed temperature and salinity rather than the predicted. Comparison of these four values with Wilson's sound velocity value for the same data is made to determine the adequacy of the regression equations.

The comparisons made are as follows: Wilson's - Mackenzie's, Wilson's - regression sound velocity, and Mackenzie's - regression sound velocity, using the observed temperature and salinity. The same comparisons are made using the predicted temperature and salinity.

A comparison is made between the general regression equation where depth is an independent variable and the case where equations for temperature, salinity, and sound velocity

are built at each depth plane.

The reliability of the Mackenzie and Wilson equations will be discussed in Chapter II and the modification to Mackenzie's equation needed to obtain agreement with Wilson's equation will be discussed in Chapter III.

Data was made available on punched cards by Mr. Richard Bolton by programs to decode the "Rapid Access Tape Format Oceanographic Station Data" system developed and provided by Mr. Walter E. Yergen.[3]

The cards consist of 3720 observations for latitude, longitude, depth, temperature, salinity, day-of-year, and Wilson's sound velocity value computed from these variables using a procedure described in Chapter II.

In order to develop more meaningful models, a decision was made to investigate a 4° by 4° square in the North Atlantic Ocean rather than several 2° by 2° squares in the same area. It was felt that if adequate prediction equations could be built for this area, then certainly the same equations would be adequate for each of the four 2° by 2° squares contained in the 4° by 4° square. Since excellent prediction equations were obtained for the 4° by 4° square, 36° - 40° north latitude and 68° - 72° west longitude, the remainder of the study was devoted to investigating 2° by 2° squares around this 4° by 4° square.

II. REVIEW OF LITERATURE

Many sound velocity tables have been developed for both distilled water and sea water. N. H. Heck and J. H. Service[4] published a set of tables in 1924, which were based on a systematic calculation scheme. In 1927, D. J. Matthews[5] published a table of sound velocity calculations for distilled water and sea water. In 1939, Matthews published a revised edition of his tables after the improved set of tables of Kuwahara[6] were introduced in 1938. The revised edition of Matthews was in close agreement with Kuwahara, but the Kuwahara tables are considered to be the better of the two.

The Kuwahara tables motivated the development, by several individuals and organizations, of equations to represent this data. Three of the better known and more reliable equations developed to represent the Kuwahara tables are those of H. V. Mackenzie, Wayne D. Wilson, and V. A. Del Grosso.[7] The Mackenzie and Wilson equations will be discussed in some detail since they are used as support in the substantiation of results in this study. Results of Del Grosso's study are used in the modification of Mackenzie's equations to reduce residuals at upper depths.

The basic Mackenzie equation of form

$$V_{TSD} = V_{0,35,0} + \Delta V_T + \Delta V_S + \Delta V_D + \Delta V_\phi + \Delta V_{TSD} \qquad (1)$$

is readily seen to be a function of Temperature (T), Salinity (S), Depth (D), and Latitude (ϕ, absolute value of).

The equation consists of 6 parts:

1. Reference velocity, $V_{0,35,0}$, computed at 0°C, 35% salinity, and zero depth.
2. Temperature dependence, (ΔV_T).
3. Salinity dependence, (ΔV_S).
4. Depth dependence, (ΔV_D).
5. Latitude dependence (ΔV_ϕ).
6. Interaction dependence due to simultaneous change of T, S, D (ΔV_{TSD}).

   ΔV_{TSD} is broken into three parts ΔV_{TS}, ΔV_{SD}, and ΔV_{TD} for further analysis where

   a. ΔV_{TS} = Temperature - salinity interaction

   b. ΔV_{SD} = Salinity - depth interaction

   c. ΔV_{TD} = Temperature - depth interaction

where

1. $V_{0,35,0} = 1445.5$ M/S (2)
2. $\Delta V_T = 4.6374\ T - 5.383\times10^{-2}\ T^2 + 2.543\times10^{-4}T^3$ (3)
3. $\Delta V_S = 1.307(S-35)$ (4)
4. $\Delta V_D = 1.815\times10^{-2}D - 5.291\times10^{-12}D^3$ (5)
5. $\Delta V_\phi = 1.5\times10^{-6}D(\phi-35) + 0.94\times10^{-12}D\ (\phi-35)^2 - 2.94\times10^{-18}D^3(\phi-35)^3 - 1.214\times10^{-3}(\phi-35)$ (6)
6. $V_{TSD} = V_{TS} + V_{SD} + V_{TD}$ (7)

where

a. $\Delta V_{TS} = (S-35)[-1.07\times10^{-2}T + (5.0\times10^{-5} - 4.1\times10^{-8}D)T^2]$

b. $\Delta V_{SD} = (S-35)(3.36\times10^{-5}D - 4.55\times10^{-9}D^2)$

c. $\Delta V_{TD} = D(-1.9\times10^{-6}T^2 + 6.35\times10^{-8}T^3 + 4.1\times10^{-10}T^4) + T(6.95\times10^{-6}D - 5.27\times10^{-9}D^2 + 2.7\times10^{-14}D^3)$

summing the results (2) - (7) give the result (1).

The Mackenzie equation agrees with the Kuwahara tables to within .1 M/sec everywhere, but it should be noted that the equations were developed to fit this particular data. This is not to say that the equations will not be useful in data reduction for other areas, but one should not be disappointed in finding larger residuals between Mackenzie's values and actual readings or between Mackenzie's values and Wilson's values for the same data.

Mackenzie's equations are flexible and provision is made for modification if necessary. There is evidence in the analysis to support the fact that the depth dependency factor and/or the latitude dependency factor need modification. Experimentation with this problem will be discussed in the following chapter.

The formulation of Wilson's equation[2] displays the same basic form as Mackenzie's equation; that is,

$$V = 1449.22 + \Delta V_T + \Delta V_P + \Delta V_S + \Delta V_{STP} \qquad (8)$$

The main differences are that V is a function of temperature, salinity and pressure, where pressure is a function of depth. The equations were developed in a controlled laboratory environment. The development was restricted in the assumption that 99.5% of all sea water falls in the ranges of $-3°C < T < 30°$ for temperature, $1.033\ kg/cm^2 < P < 1000.0\ kg/cm^2$ for pressure, and $33°/_{oo} < S < 37°/_{oo}$ for salinity. The equations were developed over 581 laboratory measured sound speeds for fifteen temperatures, eight pressures, and

five salinities. The method of least squares was applied, using a 20x20 matrix to arrive at the coefficients.

The breakdown of equation (8) is as follows:

1. 1449.22 = reference velocity computed at T=0°C, P=0.0 kg/cm^2, and S = 35°/$_{oo}$ (parts/$_{1000}$). (9)

2. Temperature contribution

$$\Delta V_T = 4.6233T - 5.4585\times10^{-2}T^2 + 2.822\times10^{-4}T^3 - 5.07\times10^{-7}T^4 \quad (10)$$

3. Pressure contribution

$$\Delta V_P = .160518P + 1.0279\times10^{-5}P^2 + 3.451\times10^{-9}P^3 - 3.503\times10^{-12}P^4 \quad (11)$$

4. Salinity contribution

$$\Delta V_S = 1.391(S-35) - 7.8\times10^{-2}(S-35)^2 \quad (12)$$

5. Interaction contribution for simultaneous changes

$$\begin{aligned}\Delta V_{STP} = {} & (S-35)[-1.197\times10^{-2}T + 2.61\times10^{-4}P \\ & - 1.96\times10^{-7}P^2 - 2.09\times10^{-6}T\,P] \\ & + P\,[-2.796\times10^{-4}T + 1.3302\times10^{-5}T^2 \\ & - 6.44\times10^{-8}T^3\,] \\ & + P^2[-2.391\times10^{-7}T^2 + 9.286\times10^{-10}T^2] \\ & - 1.745\times10^{-10}P^3T\end{aligned} \quad (13)$$

Summing the results (9)-(12) give the result (8).

In order to use equation (8) pressure must be expressed as a function of depth. Ultimately, pressure is, in fact, a function of depth, salinity, gravitational attraction, and temperature.

Wilson[2] specifies that pressure may be found by dividing depth into incremental layers and summing the product

of average density in each layer times the thickness of the layer. This is expressed as $P_i = \Sigma \bar{g}_\theta \bar{\ell}_i t$ where g_θ is the acceleration due to gravity at latitude θ and at the mean depth of the layer, $\bar{\ell}_i$ is the average density of the layer and t is the thickness of the layer.

A more complete approach for determining pressure at depth D_i is outlined by Walter Yergen.[8] The development is based on the assumption that as initial conditions, the surface pressure is equal to the mean standard atmospheric pressure of 10.1325 decibars and that the initial gravitational attraction g_o may be computed as a function of Latitude (θ) according to

$$g_o = .980616 - 2.5928 \times 10^{-3} \cos(2\theta) + 6.9 \times 10^{-6} \cos^2(2\theta) \ \frac{\text{decimeters}}{\text{cm}^2} \qquad (14)$$

and that the change in g between depths is given by

$$g_i = g_o + 1.101 \times 10^{-7}(D_i - D_{i-1}). \qquad (15)$$

Since pressure is a function of density, and density is not explicitly given, an approximated density ℓ_i at D_i is attained by successive iterations $\ell_{i_1}, \ell_{i_2}, \ldots, \ell_{in}$. In theory the iteration should stop when the difference $|\ell_{i,j+1} - \ell_{i,j}| \leq \varepsilon$ where ε is some predefined tolerance. Then ℓ_i for D_i is taken to be ℓ_{ij}.

The determination of pressure (P_i) at depth D_i is an iterative procedure of successive alternating approximations between pressure (P) and density ℓ in the sequence

$$P_{i_1}, \ell_{i_1}, P_{i_2}, \ell_{i_2}, \ldots, P_{in}, \ell_{in}.$$

This requires that an initial density ℓ_o be known, and

initial pressure, P_o which is assumed to be 10.1325 decibars. The first approximation to the true pressure at D_i is taken to be $P_{i_1}=P_{i-1}$.

Once the first density approximation, ℓ_{i_1}, is computed, then equation (16) is used to compute the first approximation to the true pressure.

$$P_{ik} = P_{i-1} + \tfrac{1}{2}(\ell_{i-1}+\ell_{i,k})\,\bar{g}_i(D_i-D_{i-1}),\ k=1. \qquad (16)$$

where $\bar{g}_i = \tfrac{1}{2}(g_{i-1}+g_i)$ and g_i is from (15).

Now that ℓ_i, ℓ_{i_1}, P_o, P_{i_1}, are known, a second approximation to the true density, ℓ_{i_2}, is computed. For ℓ_{ik} at depth D_i where $k \geq 2$, the following expression is used,

$$P_{ik} = (1+10^{-3}\sigma_t)/R \qquad (17)$$

where σ_t and R are functions of temperature, salinity, and the previously computed $P_{i,k-1}$ according to the following relations

$$\begin{aligned} 10^{-3}\sigma_t = {} & (3.118633\times10^{-6} + 4.5317157\times10^{-3}T \\ & -5.4593903\times10^{-4}T^2 -1.4385354\times10^{-10}T^4)/(67.26+T) \\ & +\sigma_o\,(.001-4.7867\times10^{-6}T +9.8185\times10^{-8}T^2 \\ & - 1.0843\times10^{-9}T^3) +\sigma_o^2\,(1.803\times10^{-6}T \\ & -8.164\times10^{-10}T^2 + 1.667\times10^{-11}T^3) \end{aligned} \qquad (18)$$

where $\sigma_o = -9.3445863\times10^{-2} +.81487658S -4.8249614\times10^{-4}S^2 +6.7678614\times10^{-6}S^3$ (19)

and $R = 1-[4.886\times10^{-6}P/(1+1.83\times10^{-5}P)]$

$$\begin{aligned} & +P[-22072\times10^{-7}+3.673\times10^{-8}T - 6.63\times10^{-10}T^2 \\ & +4\times10^{-12}T^3 + \sigma_o(1.725\times10^{-3}-3.28\times10^{-10}T \\ & +4\times10^{-12}T^2) +\sigma_o^2(-4.5\times10^{-11}+10^{-12}T)] \\ & +P^2[-6.68\times10^{-14} -1.24064\times10^{-12}T +2.14\times10^{-14}T^2 \end{aligned}$$

$$+\sigma_o(-4.248\times10^{-13}+1.206\times10^{-14}T-2\times10^{-16}T^2)$$
$$+\sigma_o^2(1.8\times10^{-15}-6\times10^{-17}T)] + P^3 1.5\times10^{-17}T \qquad (20)$$

where, for the second approximation to density $\ell_{i,2}$, P= P_{i_1}. Now find $P_{i,2}$ by using (16) with k = 2. This back and forth iteration is continued until $|P_{in}-P_{i,n-1}| \leq \varepsilon$. The data, however, is somewhat inaccurate and warrants no more than three iterations as a best approximation to the true pressure at D_i. Hence P_{i_3} is used in (10), (11), (12), (13) for finding sound velocity. Note that since the above pressure is in decibars, the conversion $P = .101971P_{i_3}$ must be made before use in Wilson's equation. If the velocity is desired in feet per second $V_{feet/sec} = V_{meters/sec}\cdot(3.28083)$ yields the desired result.

Wilson and Del Grosso concluded from careful laboratory measurements that the reference velocity, $V_{0,35,0}$, used by Kuwahara in constructing the Kuwahara tables is low by about 3 m/sec, particularly at upper depths where pressure is lower. A comparison of Wilson's predicted values and the values predicted by Kuwahara, substantiates this 3 m/s differential from 0 to 100 kg/cm^2 pressure. The reference velocity in Mackenzie's equation (8) will be low by 3 m/s also since the equation was constructed to fit the Kuwahara tables. The 3 m/sec differential in Kuwahara's values at atmospheric pressure is concluded to be a result of slightly erroneous data on the compressibility of water.[7]

Comparing the results of Mackenzie's and Wilson's equations when applied to oceanographic data, other

than that for which the equations were developed, substantiates the 3 m/s difference at near atmospheric pressures and below. Wilson's equation predicted values almost consistently 3 m/s higher than Mackenzie for this data, particularly for depths to 500 meters. Beyond this depth there is, roughly, a linear decrease in the differences of sound velocity predicted by the two equations at the same temperature, salinity, depth observation. The two equations, at 2000 meters are in excellent agreement. Figure 1 shows, roughly, the plot of residuals from depth 0-2500 meters.

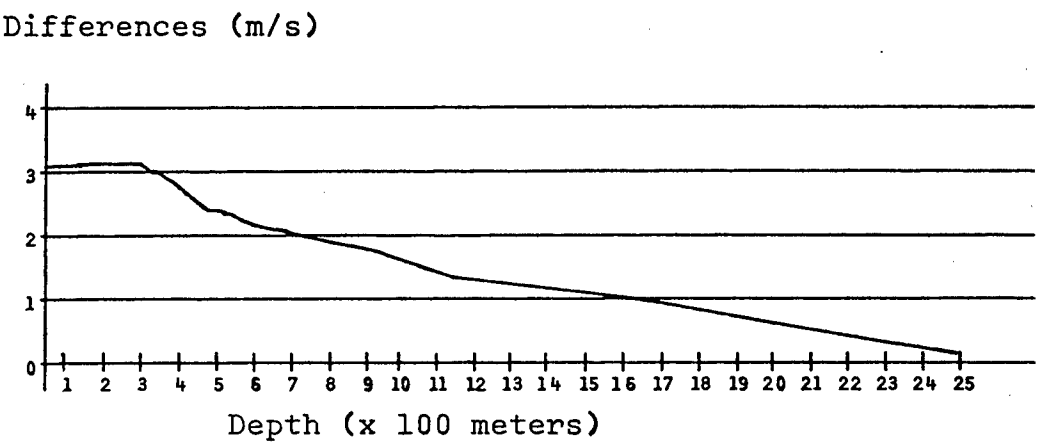


Figure 1. Plot of the residuals between Wilson's values and Mackenzie's values for the same data. ($r_i = w_i - m_i$)

The values yielded by Wilson's equations are used extensively in checking the results of this study since these values are considered to be good for most applications in the physical sciences.[2] Mackenzie's equation, however, is easier to use since there is no pressure dependency term. The modification to Mackenzie's equation, to be discussed in the following chapter, is warranted on the basis of its ease

of use and by the fact that, for the areas considered, the differences from Wilson's values were no greater in absolute value than .9 meters/second for all depths. Both equations are used, however, in checking the results of the regression equations developed in this investigation.

Data deficiencies are always an impediment in the solving of oceanographic problems. C. J. VanVliet has made a rather extensive empirical study on the effect of random and nonrandom missing data on regression and autocorrelation analyses of time series data.[10] The time series analysis is to isolate trend or a gradual increase or decrease in a system over a long period of time, oscillation or a variation about the trend which occurs with a pattern of regularity over a period of time, and random elements or unpredictable variations in a given variable.

Van Vliet considered the surface temperature variable in his analysis. The Monte Carlo method was employed to simulate missing data situations, random and nonrandom. The regression and autocorrelation coefficients were computed for each time series analysis.

A determination of the sensitivity of coefficient variability due to random and nonrandom missing data was made for different series lengths. The conclusion was that if the missing data is random, a smaller sample size is used, and the change in the variability of the regression coefficients is predictable by the amount of reduction in sample size. The random deletion of data increases both

regression and autocorrelation coefficient variability.

For nonrandom missing data, or an excessive number of longer sequences of missing data, the increase in the variance of the regression coefficients is roughly twice the increase for random missing data. For nonrandom missing data the increase in variance of the autocorrelation coefficients is roughly 1.2 times the increase attributable to random missing data. The above suggests that the autocorrelation coefficients are less sensitive to the effects of nonrandom missing data than the regression coefficients.

E. R. Anderson, using regression and autocorrelation techniques, determined that in order to eliminate short term variability and reliably estimate sea-surface temperature, a time series record of 8 to 10 years is needed.[11] Anderson developed a regression model considering latitude, longitude, and day of year as independent variables.[9] This model was found capable of estimating seasonal variation of sea-surface temperature off the west coast of the United States, in water depths of greater than 100 fathoms, to a standard deviation of less than 1°F. Anderson's model: $T_s = F$ (Latitude, Longitude, day-of-year).

$$T_s = \beta_0 + \beta_1 D_y + \beta_2 D_y^2 + \beta_3 D_y^3 + \beta_4 D_y^4 + \beta_5 D_y^5 \quad \text{(day-of-year)}$$

$$+\beta_6 L_a + \beta_7 L_a^2 + \beta_8 L_a^3 \quad \text{(latitude)}$$

$$+\beta_9 L_o + \beta_{10} L_o^2 + \beta_{11} L_o^3 \quad \text{(longitude)}$$

$$+\beta_{12} L_a D_y + \beta_{13} L_a D_y^3 + L_a D_y^5 \quad \text{(latitude-day)}$$

$$+\beta_{15} L_o D_y + \beta_{16} L_o D_y^3 + \beta_{17} L_o D_y^3 \quad \text{(longitude-day)}$$

$$+ \beta_{18} L_a L_o + \beta_{19} L_a L_o^2 + \beta_{20} L_a L_o^3 \qquad \text{(latitude-longitude)}$$
$$+ \beta_{21} L_a^2 L_o + \beta_{22} L_a^3 L_o$$

where L_a = latitude

L_o = longitude

D_y = Day-of-year

It should be pointed out that the present study is primarily a search for adequate models to represent temperature, salinity, and sound velocity and the data used is primarily from one area and one season. The seasonal variation, therefore, will not be as pronounced as in Anderson's study. The terms of Anderson's model, however, are incorporated into one of the more complex temperature models to be discussed in Chapter 3. This model is also expanded to include depth as an independent variable. It is hoped that this technique will help in explaining variability of temperature to depths of 500 meters.

III. DISCUSSION

The extremely dynamic character of the ocean environment is a formidable obstacle in the search for stable techniques for predicting ocean variables. The oceanographic problem, then, becomes one of searching for "adequate" models to use in reduction of available data.

This chapter presents the results of a preliminary inquiry into the feasibility of eliminating the need for "on-location" measurements of temperature and salinity by building multiple regression models to predict these variables as functions of geographic location, time, depth, and day-of-year.

The regression development consists of a systematic consideration of polynomial models of the form

$$Y = \sum_{i=1}^{N} \beta_i X_i + \varepsilon$$

where

$$X_i = f(Z_1^{\alpha_1}, Z_2^{\alpha_2}, \ldots, Z_n^{\alpha_n})$$

such that: Z_i are independent variables,

and: α_i are powers of the independent variables

and: ε is the error.

The models tried vary in complexity, from second order models with only two independent variables (latitude and longitude), to tenth order models with 6 independent variables (latitude, longitude, depth, temperature, salinity, day of year). The investigation proceeded from producing models for individual depth planes to a general regression

situation for which depth was an independent variable. The final, more involved model, for temperature, contains the terms of the model described by E. R. Anderson[9] to account for seasonal variation in temperature.

Mackenzie's equation proves to be a valuable tool for comparing regression results with some existing standard; however, Figure 1 reveals a deficiency in predicting particularly at depths to 1500 meters. The nearly linear decrease in the magnitude of the residuals, $r_i = w_i - m_i$, where w_i is Wilson's value and m_i is Mackenzie's value at observation i, suggests a slight modification in the depth dependency term is in order. The reference velocity is taken as that of Del Grosso[1], $V_{0,35,0} = 1448.5$, and an amount .0012D is subtracted from the depth dependency term. That is, now:

$$\Delta V_D = 1.815\times10^{-2}D - 5.291\times10^{-12}D^3 - 1.2\times10^{-3}D = 1.63\times10^{-2}D - 5.291\times10^{-12}D^3$$

Notice that at upper depths the change in the depth dependency term is negligible, but since the reference velocity is 3 m/s greater, Mackenzie's equation predicts very close to Wilson's. As depth increases, the depth dependency change becomes more pronounced, until at 2500 meters the effect of the higher reference velocity is cancelled (i.e., .0012(2500)=3), and Mackenzie's equation is predicting as it was originally.

Figure 2 shows a plot of the residuals after modification of Mackenzie's equation and Figure 3 shows the distribution of residuals by means of a histogram, after

this modification.

The magnitude of 3718 of the 3720 residuals obtained were of the order $|r_i| \leq 0.9$. The two residuals whose value was greater than 1.0, were found at a zero salinity reading. Clearly, for this area, Mackenzie's equation is much improved, and will be very beneficial for comparing to regression sound velocity predictions.

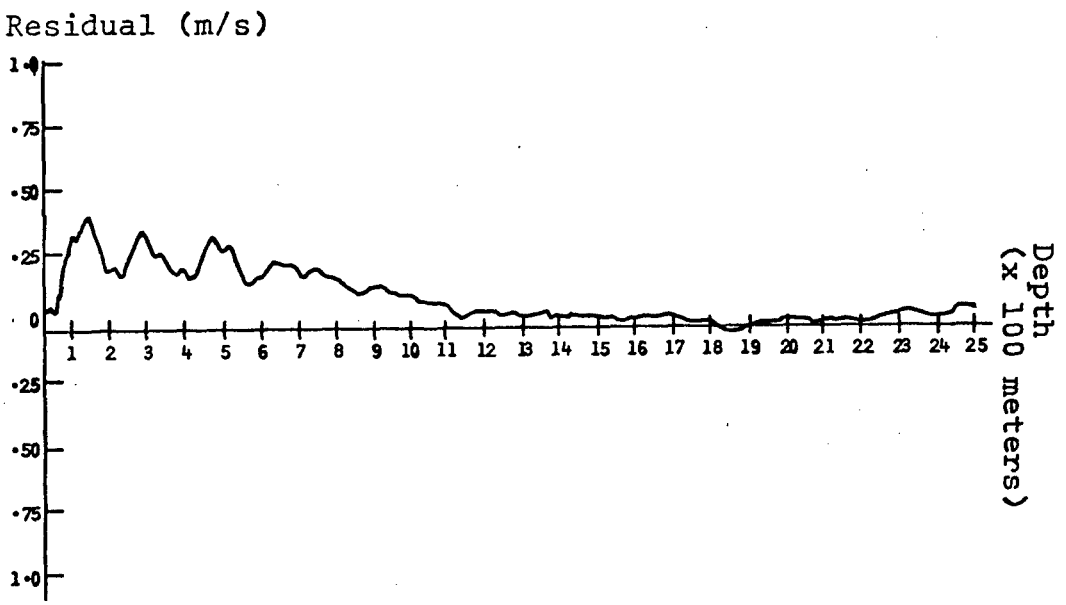
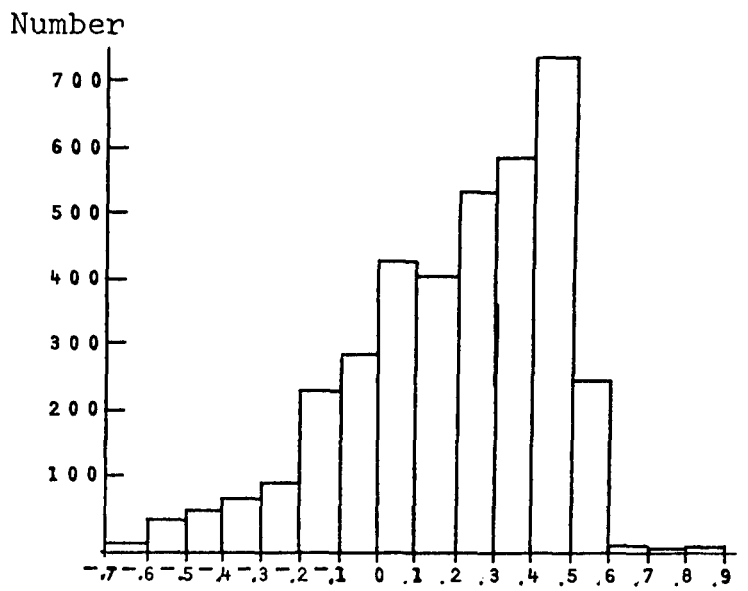


Figure 2. Residuals Versus Depth After Modification of Mackenzie's Equation ($r_i = w_i - m_i$)

Figure 2 represents a plot of the average residual ($r_i = w_i - m_i$) at depth D_i. The plot does not show the residuals which reached larger values (e.g., $\geq .5$). For this reason the distribution is shown in Figure 3 as a histogram. The plot is shown as number of residuals against magnitude of residual. For example, the number of residuals from 0.0 to 0.1 is 428. Alternating positive and negative residuals lower the value of the average r_i at D_i in Figure 2.

After modification, Mackenzie's equation was checked on a 1° by 1° square of data from 38°-39° N latitude and 69°-70° W longitude. This run substantiated the validity of the modification, for all the residuals ($w_i - m_i$) here fell in the range of -.5 m/s to .9 m/s.



From	To	Number
-.7	-.6	5
-.6	-.5	38
-.5	-.4	44
-.4	-.3	63
-.3	-.2	96
-.2	-.1	221
-.1	0	283
0	.1	428
.1	.2	403
.2	.3	531
.3	.4	597
.4	.5	738
.5	.6	259
.6	.7	5
.7	.8	3
.8	.9	4
Total		3718

Figure 3. Histogram of residual distribution of $r_i = w_i - m_i$ for the same data in the 4°x4° square 36°-40° N latitude 68°-72° W longitude.

Throughout the remainder of this discussion, w_i and m_i will represent Wilson's and Mackenzie's sound velocity, respectively, as before, and B_i will represent the sound velocity yielded by the regression equation.

The stepwise multiple regression procedure was utilized in building polynomial models involving two to six independent variables and various higher order cross products of these variables in ascending order of complexity. The greatest significance is attached to the more complex models

toward the end of the study and therefore the most comprehensive analysis is reserved for those models described on pages 31 - 34.

The set of data used is from the 4^{o} by 4^{o} square 38^{o} - 40^{o} N latitude, 68^{o} - 72^{o} W longitude, consisting of 3720 data points over 20 depth planes of 0, 10, 20, 30, 50, 75, 100, 150, 200, 250, 300, 400, 500, 600, 800, 1000, 1200, 1500, 2000, 2500 meters.

Some arbitrary criterion must be established for measuring how well the regression equations appear to be in the analysis. This may be achieved in several ways. This investigator will use three common criterion for determining goodness of fit. First, and probably most important, is the R^2 ratio or percent of variation explained by the regression equation; second, the standard error of the regression equation; and third, plots of the residuals (deviation from actual value) against the dependent variable ($\hat{y}$). Ideally, we wish to increase R^2 as we decrease the standard error of $\hat{y}$.

The stepwise procedure requires a significance level for the deletion of non-significant terms from the model and the addition of significant terms. In most of the ensuing models, an F level of 2.65 is used for adding and deleting variables in the model building process. This figure represents $F(1,v_2,.90)$ where $v_2 \geq 120$ degrees of freedom.

When plotting the residuals $(y_i - \hat{y}_i)$ against $\hat{y}$, four common patterns may appear signifying certain conditions of the prediction equation over the range of the dependent

variable. Figure 4 shows the general shape of these patterns. Variations of shape, slope, and combinations of more than one are possible.

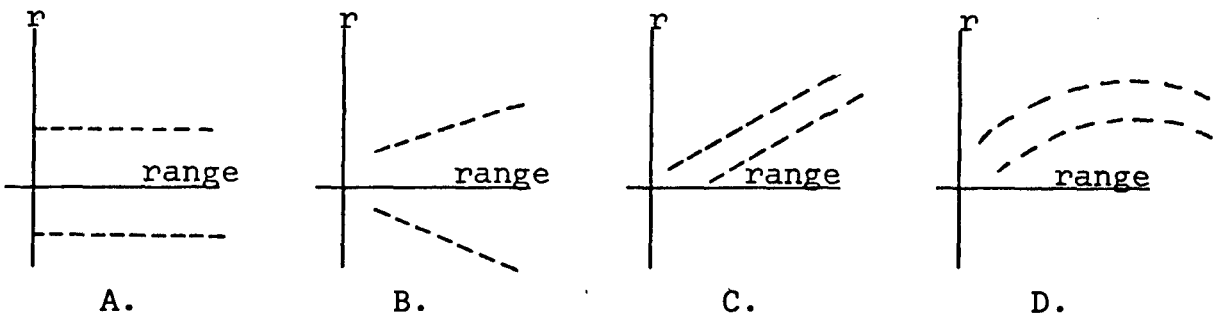


Figure 4. Possible patterns of residual plot of $y_i - \hat{y}_i$ against $\hat{y}$.

Interpolation of the cases is as follows:[12]

A. Residuals fall in a horizontal band indicate no unaccounted for effects over the range of the dependent variable $\hat{y}$. This indicates a normal regression situation and good fit.

B. Residual plot forms a fan pattern indicating the variance is not constant but increases with increasing values of the dependent variable. This implies weighted least squares analysis should be used instead.

C. Band with slope greater than zero indicating that a linear term is needed in the model.

D. Nonlinear band indicates linear and quadratic terms are needed in the model.

This type of analysis will be applied to more significant models.

Second Order Model - Two Independent Variables

The second order model was the simplest of all models tried. The purpose was to determine if temperature, salinity, and sound velocity are functions of geographic location (latitude and longitude). Temperature, salinity and sound velocity are used as independent variables. Depth is not an independent variable here, consequently the model is applied to the data at each depth plane for each dependent variable.

Model 1 used here is as follows:

$$\beta_0+\beta_1 Z_1+\beta_2 Z_1^2+\beta_3 Z_2+\beta_4 Z_1 Z_2+\beta_5 Z_1^2 Z_2+\beta_6 Z_2^2+\beta_7 Z_1 Z_2^2+\beta_8 Z_1^2 Z_2^2$$

where: Z_1 = latitude

Z_2 = longitude

Figure 5 shows the R^2 and corresponding standard error σ_T of $\hat{T}$ for each depth plane when temperature is the dependent variable. Figures 6 and 7 show plots of the R^2 statistic and corresponding standard error, for each depth plane, where salinity and sound velocity are the dependent variables, respectively.

An examination of the residuals from the resulting equations and the plots in figures 5, 6, and 7 reveal deficiencies in Model 1. The residuals (actual - predicted) are generally in the ranges of $\pm 5^{o}C$ for temperature, $\pm 8^{o}/_{oo}$ for salinity, and ± 50 meters per second for sound velocity. These residuals are too large in comparison to the magnitude of numbers being predicted and indicate an obvious need for more independent variables and/or higher order cross products in the model.

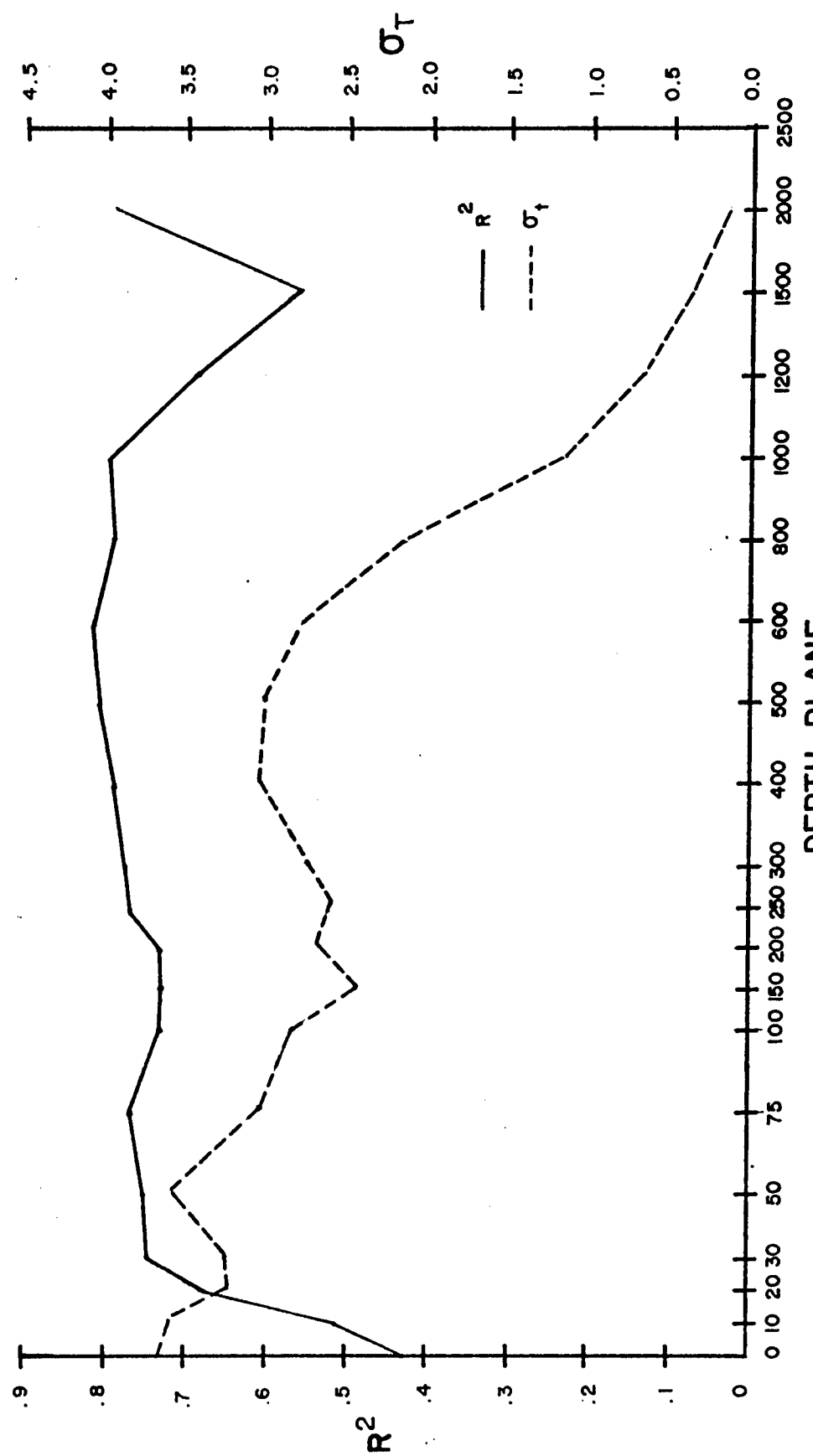


FIGURE 5 R^2 VERSUS σ_T FOR MODEL I

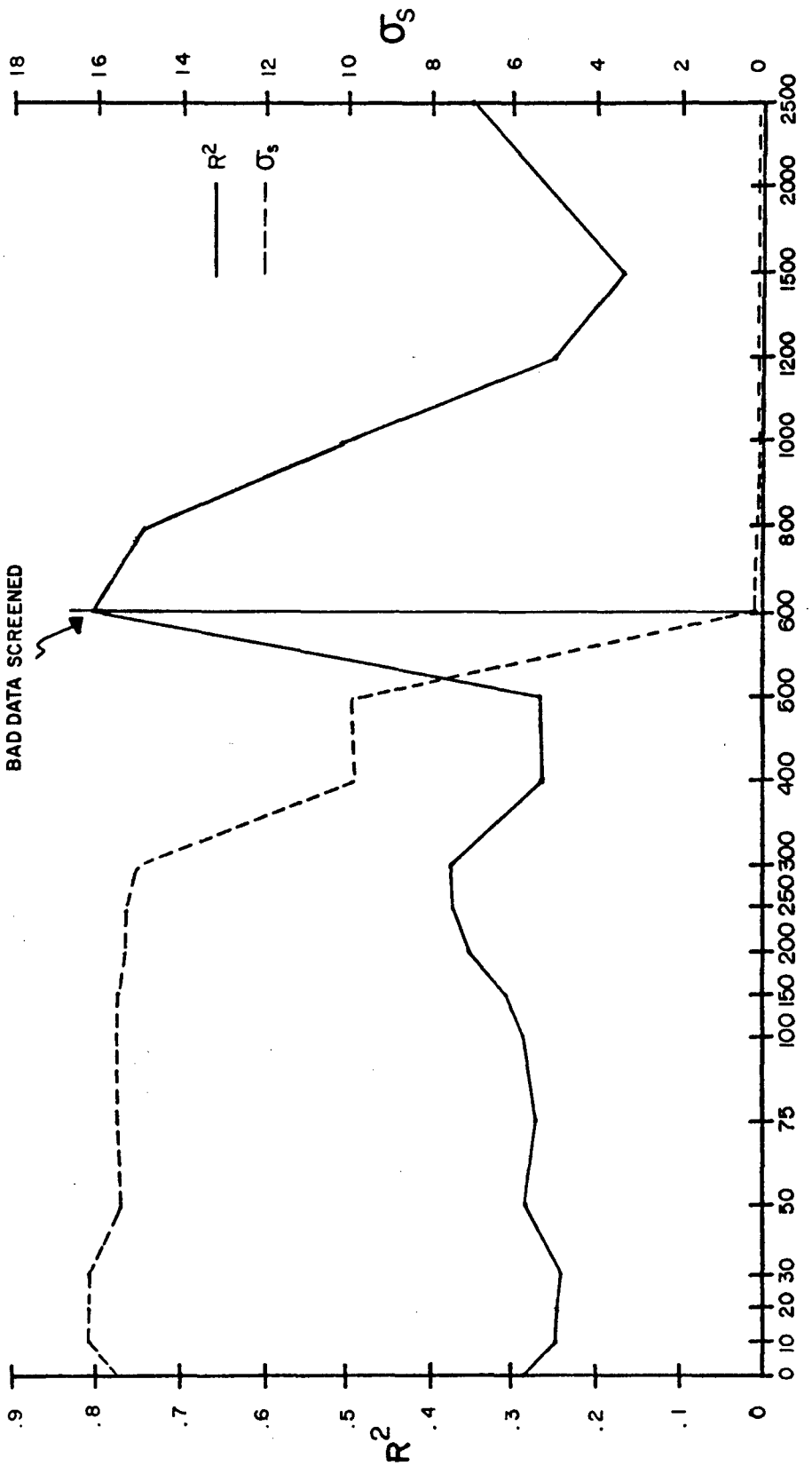


FIGURE 6. R^2 VERSUS σ_s FOR MODEL I

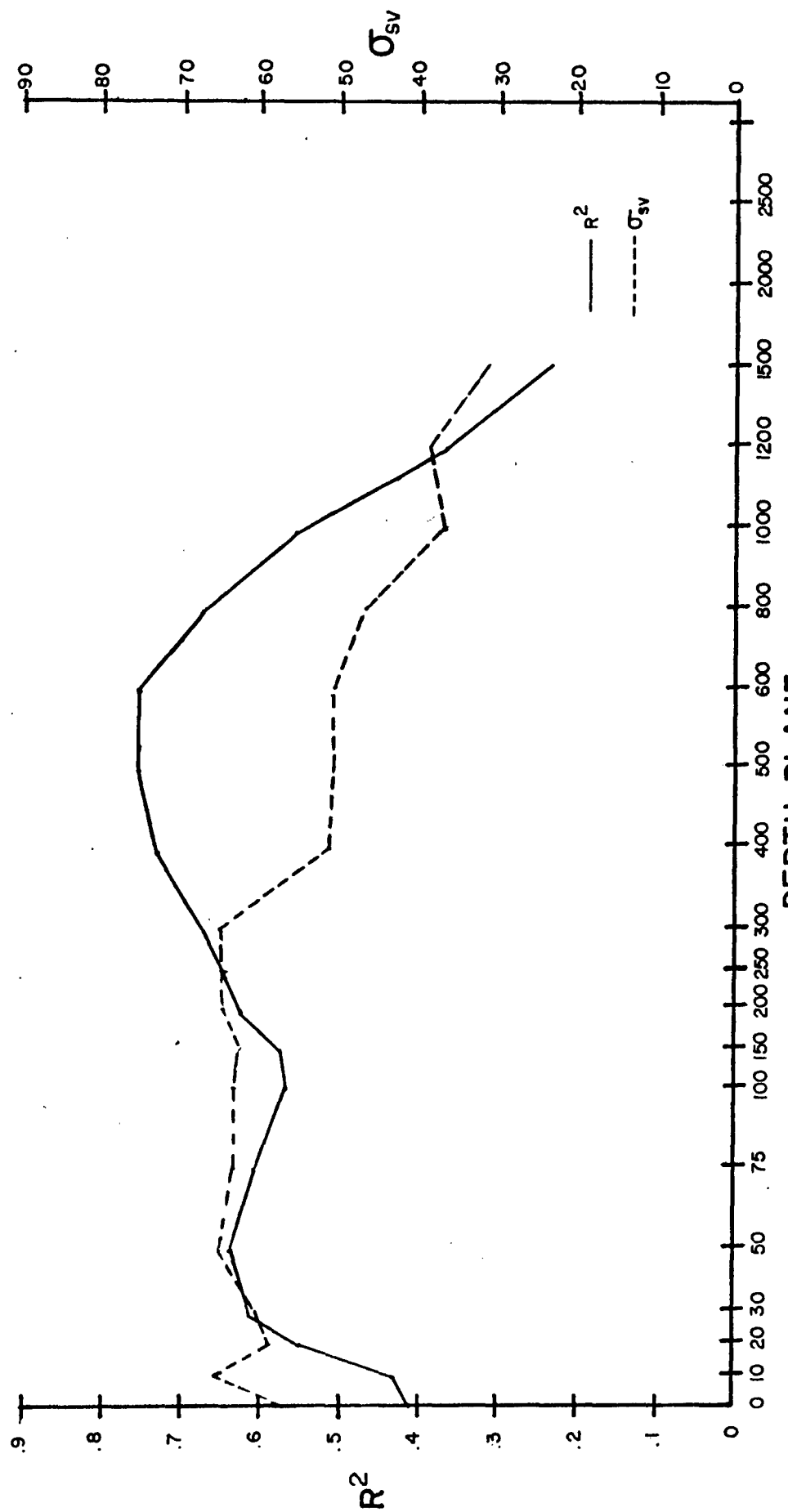


FIGURE 7. R^2 VERSUS σ_w FOR MODEL I

A further check was performed on the results of Model 1 by evaluating Mackenzie's equation using the temperature and salinity yielded by the regression equations rather than the true temperature and salinity. The sound velocity obtained by Mackenzie's equation in this manner was then compared to Wilson's sound velocity value and then to the sound velocity predicted by the regression sound velocity equation for the same data. Mackenzie's equation utilizing the calculated temperature and salinity displayed severe differences from Wilson values. In many cases the differences were 800 m/sec! The differences between Mackenzie's prediction and the regression equation prediction were even more severe. Some of the differences here reached 900 m/sec. The inadequacy of Model 1 was substantiated, and a more expanded model was tried.

Much of the inadequacy of Model 1 and the rather wild results obtained in the analysis is attributable to missing data resulting from such things as instrument failure or bad weather. To eliminate as much of the effect of missing data as possible, a screening is implemented so that if a zero temperature or salinity reading is encountered, it is essentially eliminated from the discussion. Figure 6 shows the effect of screening out bad data.

Second Order Model - Four Independent Variables

In this facet of the study the regression model was expanded to include four independent variables, latitude,

longitude, day of year, and time of day.

The introduction of additional independent variables greatly increases the possible combinations of cross products which could be considered to enter the model. A judicious choice was made and the resulting model was:

$$\beta_0+\beta_1 Z_1+\beta_2 Z_1^2+\beta_3 Z_1 Z_2+\beta_4 Z_1 Z_3+\beta_5 Z_1 Z_4+\beta_6 Z_2+\beta_7 Z_2^2$$
$$+\beta_8 Z_2 Z_3+\beta_9 Z_2 Z_4+\beta_{10} Z_3+\beta_{11} Z_3^2+\beta_{12} Z_3 Z_4+\beta_{13} Z_4+\beta_{15} Z_4^2$$
$$+\beta_{16} Z_1 Z_2 Z_4+\beta_{17} Z_1 Z_3 Z_4+\beta_{18} Z_2 Z_3 Z_4+\beta_{19} Z_1 Z_2 Z_3 Z_4+\varepsilon \quad \text{(Model 2)}$$

where Z_1 = day-of-year, Z_2 = time-of-day, Z_3 = latitude, Z_4 = longitude.

Higher order terms were arbitrarily avoided at this point to minimize the complexity of the problem in the early stages. Notice, however, Z_1^2, Z_2^2, Z_3^2, Z_4^2 have been included.

Seven depth planes were chosen for the analysis; 0, 10, 20, 50, 100, 500, 1500 meters. Model 2 was applied to the data at each depth plane for each of the dependent variables temperature, salinity, and sound velocity. $\hat{T}$, $\hat{S}$, $\hat{SV}$ were determined at each depth plane with the 90% F of 2.65. Table I shows the R^2 statistic and corresponding standard error for each regression equation at each depth plane considered.

Depth in meters	R^2 $(\hat{T})$	σ_t	R^2 $(\hat{S})$	σ_s	R^2 $(\hat{SV})$	σ_{sv}
0	.666	3.07	.378	13.09	.596	53.1
10	.721	2.92	.346	13.75	.512	62.9
20	.813	2.56	.736	.61	.627	53.9
50	.806	3.15	.705	.62	.640	64.5
100	.774	2.57	.751	.37	.627	60.4
500	.812	2.98	.813	.35	.756	50.9
1500	.552	.39	.185	.073	.265	31.4

Table I. The R^2 and corresponding standard error for all equations developed using Model 2 at depth planes indicated.

The residuals associated with the regression equations at the various depth planes still showed excessively large deviations from the observed values. Residual patterns were similar to those of Model 1. The residuals for the equations derived from Model 2 still were generally in the range of $\pm 5°C$ for temperature, $\pm 8°/_{oo}$ for salinity, and ± 50m/sec for sound velocity. Calculation of Mackenzie's equation using the calculated temperature and salinity and comparing to Wilson's and the regression sound velocity for the same data showed no significant improvement over results from Model 1. Further comment on this particular model is deferred until more comprehensive models have been discussed.

In the general regression situation it was desired to create a model which involves as many significant independent variables and cross products as possible while at the

same time containing as few terms as possible to do a responsible job of predicting.

Depth should have a significant relationship to salinity and sound velocity. This introduces the general problem of developing regression equations for temperature, salinity, and sound velocity over all depth planes.

A reassessment of the basic problem reveals two unanswered questions. First, is it possible to develop regression equations to adequately predict temperature and salinity values, which could then be used in an existing equation, such as Wilson's or Mackenzie's equation, to yield a sound velocity value near the true value without the need for "on-location" measurements of temperature and salinity? Second, if adequate temperature and salinity equations can be developed, could a regression equation for sound velocity then be used, utilizing these predicted values, to produce sound velocities close to the true reading without relying on existing methods such as Mackenzie's or Wilson's equation? The most important aspect in either case is eliminating the need for actual measurement by instruments.

There are at least two procedures which may be used in developing the desired regression equations for temperature, salinity and sound velocity. First, one large model may be used, changing only the dependent variable. Second, an individual model for each dependent variable may be used.

It was concluded, after extensive model testing, too voluminous to present here, that the individual character

of the dependent variables temperature, salinity and sound velocity require individual models.

Thus the models presented in the ensuing discussion were built according to procedure two, and yield better results than those models tested by procedure one. It should also be pointed out that the temperature, salinity and sound velocity models presented in the following discussion are the culmination of an extensive trial and error model building process. These are the models which produced the most significant results.

$$S = \text{Salinity} = F(\text{latitude, longitude, depth})$$

$$\begin{aligned} &= \beta_0+\beta_1 Z_1+\beta_2 Z_1^2+\beta_3 Z_1 Z_1+\beta_3 Z_1 Z_2+\beta_4 Z_1 Z_3+\beta_5 Z_2 \\ &+\beta_6 Z_2^2+\beta_7 Z_2 Z_3+\beta_8 Z_3+\beta_9 Z_3^2+\beta_{10} Z_1^3+\beta_{11} Z_2^3+\beta_{12} Z_3^3 \\ &+\beta_{13} Z_1^4+\beta_{14} Z_2^4+\beta_{15} Z_3^4+\beta_{16} Z_1^5+\beta_{17} Z_2^5+\beta_{18} Z_3^5+\beta_{19} Z_1^8 \quad \text{(model 3)} \\ &+\beta_{20} Z_1^4 Z_2^4+\beta_{21} Z_1^4 Z_3^4+\beta_{22} Z_2^8+\beta_{23} Z_2^4 Z_3^4+\beta_{24} Z_3^8+\beta_{25} Z_1^3 Z_3^2 \\ &+\beta_{26} Z_2^3 Z_3^2+\beta_{27} Z_1^6+\beta_{28} Z_1^3 Z_2^2+\beta_{29} Z_2^6+\beta_{30} Z_1^{10}+\beta_{31} Z_1^5 Z_2^5 \\ &+\beta_{32} Z_1^5 Z_3^5+\beta_{33} Z_2^{10}+\beta_{34} Z_2^5 Z_3^5+\beta_{35} Z_3^{10}+\beta_{36}/Z_3+\beta_{37}/Z_3^2 \\ &+\beta_{38}/Z_3^3+\varepsilon \end{aligned}$$

where Z_1 = latitude, Z_2 = longitude, Z_3 = depth. Application of this model to the available data yielded the following prediction equation.

$$\begin{aligned} \hat{S} = &-1.096Z_1+.1635\times10^{-4}Z_1Z_3-1\times10^{-5}Z_3^2+.253\times10^{-8}Z_3^3 \\ &+.959\times10^{-5}Z_1^4+.614\times10^{-10}Z_1^3Z_3^2-.848\times10^{-11}Z_2^3Z_3^2 \\ &+.976\times10^{-9}Z_1^3Z_2^3+.62\times10^{-10}Z_2^6-.196\times10^{-23}Z_1^5Z_3^5 \\ &+.733\times10^{-25}Z_2^5Z_3^5+.8\times10^{-33}Z_3^{10}-9.6133/Z_3+8.6/Z_3^3+68.57 \end{aligned}$$

The prediction equation represents a relatively good statistical fit to the salinity observations on cards.

Nearly all residuals ($S_i - \hat{S}_i$) fall in the range -1.5 to 1.5, and R^2 = 64.14 with the standard error of $\hat{S}$ = .7375.

The temperature model is more complex since it involves two more independent variables than the salinity model, and incorporates the terms of Anderson's model[9] to account for seasonal variation. The model is expanded to include depth and day of year as independent variables.

T = Temperature = F (latitude, longitude, depth, salinity, day of year)

$$
\begin{aligned}
&= \beta_0+\beta_1 Z_1+\beta_2 Z_1^2+\beta_3 Z_1 Z_2+\beta_4 Z_1 Z_3+\beta_5 Z_1 Z_4+\beta_6 Z_1 Z_5 \\
&+\beta_7 Z_2+\beta_8 Z_2^2+\beta_9 Z_2 Z_3+\beta_{10} Z_2 Z_4+\beta_{11} Z_2 Z_5+\beta_{12} Z_3 \\
&+\beta_{13} Z_3^2+\beta_{14} Z_3 Z_4+\beta_{15} Z_3 Z_5+\beta_{16} Z_4+\beta_{17} Z_4^2+\beta_{18} Z_4 Z_5 \quad \text{(Model 4)} \\
&+\beta_{19} Z_5+\beta_{20} Z_5^2+\beta_{21} Z_1^3+\beta_{22} Z_2^3+\beta_{23} Z_3^3+\beta_{24} Z_4^3+\beta_{25} Z_5^3 \\
&+\beta_{26} e^{Z_4}+\beta_{27} Z_1^3 Z_2+\beta_{28} Z_3^4+\beta_{29} Z_4^4+\beta_{30} Z_5^4+\beta_{31} Z_1 Z_5^3 \\
&+\beta_{32} Z_2 Z_5^3+\beta_{33} Z_1 Z_2^3+\beta_{34} Z_4^5+\beta_{35} Z_5^5+\beta_{36} Z_1 Z_5^5+\beta_{37} Z_2 Z_5^5 \\
&+\beta_{38} \ln(Z_3)+\beta_{39} e^{Z_1}+\beta_{40} Z_1 Z_2^2+\beta_{41} /Z_4^2+\beta_{42} Z_1^2 Z_2 \\
&+\beta_{43} Z_4/Z_3+\varepsilon
\end{aligned}
$$

where Z_1 = latitude

Z_2 = longitude

Z_3 = depth

Z_4 = salinity

Z_5 = day-of-year

ε = error term

Notice that some experimental cross products are included in the model. It is interesting to note that some of these odd terms entered the resulting regression equation at high levels of significance. Applying this model to the

data resulted in the following prediction equation for temperature.

$$\hat{T} = -34.96 + .19\text{x}10^{-3}Z_1Z_3 + .246\text{x}10^{-2}Z_1Z_5 - .018Z_2Z_4$$
$$-.162\text{x}10^{-4}Z_3^2 + .477\text{x}10^{-3}Z_3Z_4 - .3599\text{x}10^{-4}Z_3Z_5$$
$$-.884\text{x}10^{-2}Z_4Z_5 + .492\text{x}10^{-2}Z_1^3 - .187\text{x}10^{-3}Z_2^3$$
$$+.78\text{x}10^{-8}Z_3^3 + .1466\text{x}10^{-14}e^{Z_4} + .746\text{x}10^{-4}Z_1^3Z_2$$
$$-.13\text{x}10^{-11}Z_3^4 - .513\text{x}10^{-7}Z_2Z_5^3 - .584\text{x}10^{-12}Z_1Z_5^5$$
$$+ 5.62\ \ln(Z_3) - .423Z_4/Z_3$$

This equation represents a good fit to the 3720 temperature observations on cards. For this set of data, $R^2=.9484$ and the standard error of $\hat{T}$ = 2.32. The vast majority of residuals $(T_i - \hat{T}_i)$ fall in the range $\pm 2°C$ from the observed value.

Finally, the sound velocity model used to fit the 3720 sound velocity observations is a function of five independent variables.

SV = sound velocity = F (latitude, longitude, depth, temperature, salinity)

$$= \beta_0+\beta_1Z_1+\beta_2Z_1^2+\beta_3Z_1Z_2+\beta_4Z_1Z_3+\beta_5Z_1Z_4+\beta_6Z_1Z_5$$
$$+\beta_7Z_2+\beta_8Z_2^2+\beta_9Z_2Z_3+\beta_{10}Z_2Z_4+\beta_{11}Z_2Z_5+\beta_{12}Z_3$$
$$+\beta_{13}Z_3^2+\beta_{14}Z_3Z_4+\beta_{15}Z_3Z_5+\beta_{16}Z_4+\beta_{17}Z_4^2+\beta_{18}Z_4Z_5$$
$$+\beta_{19}Z_5+\beta_{20}Z_5^2+\beta_{21}Z_1Z_2Z_3+\beta_{22}Z_1Z_2Z_4+\beta_{23}Z_1Z_2Z_5$$
$$+\beta_{24}Z_1Z_3Z_4+\beta_{25}Z_1Z_3Z_5+\beta_{26}Z_1Z_4Z_5+\beta_{27}Z_2Z_3Z_4 \qquad \text{(Model 5)}$$
$$+\beta_{28}Z_2Z_3Z_5+\beta_{29}Z_2Z_4Z_5+\beta_{30}Z_3Z_4Z_5+\beta_{31}Z_1Z_2^2Z_3$$
$$+\beta_{32}Z_1Z_2Z_4Z_5+\beta_{33}Z_1Z_3Z_4Z_5+\beta_{34}Z_2Z_3Z_4Z_5$$
$$+\beta_{35}Z_1Z_3Z_4Z_5^2+\varepsilon$$

where Z_1 = latitude

Z_2 = longitude

Z_3 = depth

Z_4 = temperature

Z_5 = salinity

and temperature will be a function of day-of-year.

Applying this sound velocity model to the available data yielded the following prediction equation:

$$\hat{SV} = (4894.08 + .0222Z_1^2 + .112Z_1Z_2 - .1127Z_3 + .373\times10^{-5}Z_3^2 + .65\times10^{-3}Z_3Z_5 - .103Z_4^2 + 3.58Z_5 + .015Z_5^2 - .0052Z_1Z_2Z_4 + .59\times10^{-4}Z_1Z_3Z_4 + .685\times10^{-6}Z_1Z_2Z_3 + .799\times10^{-5}Z_1Z_2Z_4Z_5 - .226\times10^{-7}Z_1Z_3Z_4Z_5^2)/3.281$$

This sound velocity equation is a very good fit to the data with $R^2 = .9935$ and 98% of the residuals $(SV_i - \hat{SV}_i)$ fall in the range of ± 2 m/sec. The standard error of $\hat{SV}$ = 2.9 m/sec.

The method by which these equations were derived presents an interesting possibility. A sound velocity value could be computed knowing only latitude, longitude, depth, and day-of-year, since

Salinity = F(lat, lon, depth)

Temperature = F(lat, lon, depth, salinity, day-of-year)

Sound velocity = F(lat, lon, depth, temperature, salinity)

There are now five sound velocity values for each latitude, longitude and depth.

1. Wilson's value (given in initial data)
2. Mackenzie's value computed using the observed temperature and salinity.

3. The regression equation value ($\hat{SV}$) using the observed temperature and salinity.
4. Mackenzie's value computed using the predicted temperature and salinity.
5. The regression equation value using the predicted temperature and salinity.

For each of the 3720 latitude, longitude, and depth observations, these five sound velocity values were obtained. With these five sound velocities, six comparisons were made for each data point.

1. $r_{wm} = w_i - m_i$ (Wilson's - Mackenzie's) } using observed temperature and salinity
2. $r_{wB} = w_i - B_i$ (Wilson's - Regression S.V.) } using observed temperature and salinity
3. $r_{mB} = m_i - B_i$ (Mackenzie's - Regression S.V.) } using observed temperature and salinity
4. $r_{wm} = w_i - m_i$ } using the predicted temperature and salinity
5. $r_{wB} = w_i - B_i$ } using the predicted temperature and salinity
6. $r_{mB} = m_i - B_i$ } using the predicted temperature and salinity

Six corresponding residual distributions were developed according to the magnitude of the residual. The purpose of the distributions is to determine how many of the residuals are more than 30 m/sec high, 29-30 m/sec high, . . ., 29-30 m/sec low, more than 30 m/sec low. Table II shows the six residual distributions and their densities.

Using the observed temperature (T) and salinity (S), Wilson and Mackenzie show hardly any difference as would be expected after modification of Mackenzie's equation.

Using the observed (actual) T and S the residual distribution for $w_i - B_i$ shows 98% of the residuals are in the

range $\pm$2 m/sec. This indicates a good fit to the Wilson values.

The third distribution, M - B, using the observed T and S, indicates that the regression equation is a close duplication of Mackenzie's equation; that is, only 81 of 3720 predictions differ by more than $\pm$2 m/sec. This is an interesting point, for the regression equation for sound velocity is much simpler in form than Mackenzie's equation.

The fourth distribution is obtained by comparing the Wilson sound velocity values with the Mackenzie values computed from a predicted T and S. The resulting residual distribution takes on the shape of a normal distribution, which is slightly skewed to the left. Figure 8 shows the distribution by means of histogram of magnitude against number. It is felt that the resulting distribution enhances the feasibility of predicting sound velocity given only latitude, longitude, and depth, and be at least 70% sure of being within 9 meters/sec of the true sound velocity.

The fifth residual distribution of Table II is obtained by evaluating the regression sound velocity equation using the predicted temperature and salinity and comparing the results with Wilson's value from the card (i.e., obtain all w_i-B_i). The residual distribution here is almost identical with distribution 4. The histogram of figure 8 adequately represents distribution 5 as well as distribution 4.

Distribution 6 compares the sound velocity predictions of Mackenzie's sound velocity equation to those of the

Residual Distributions

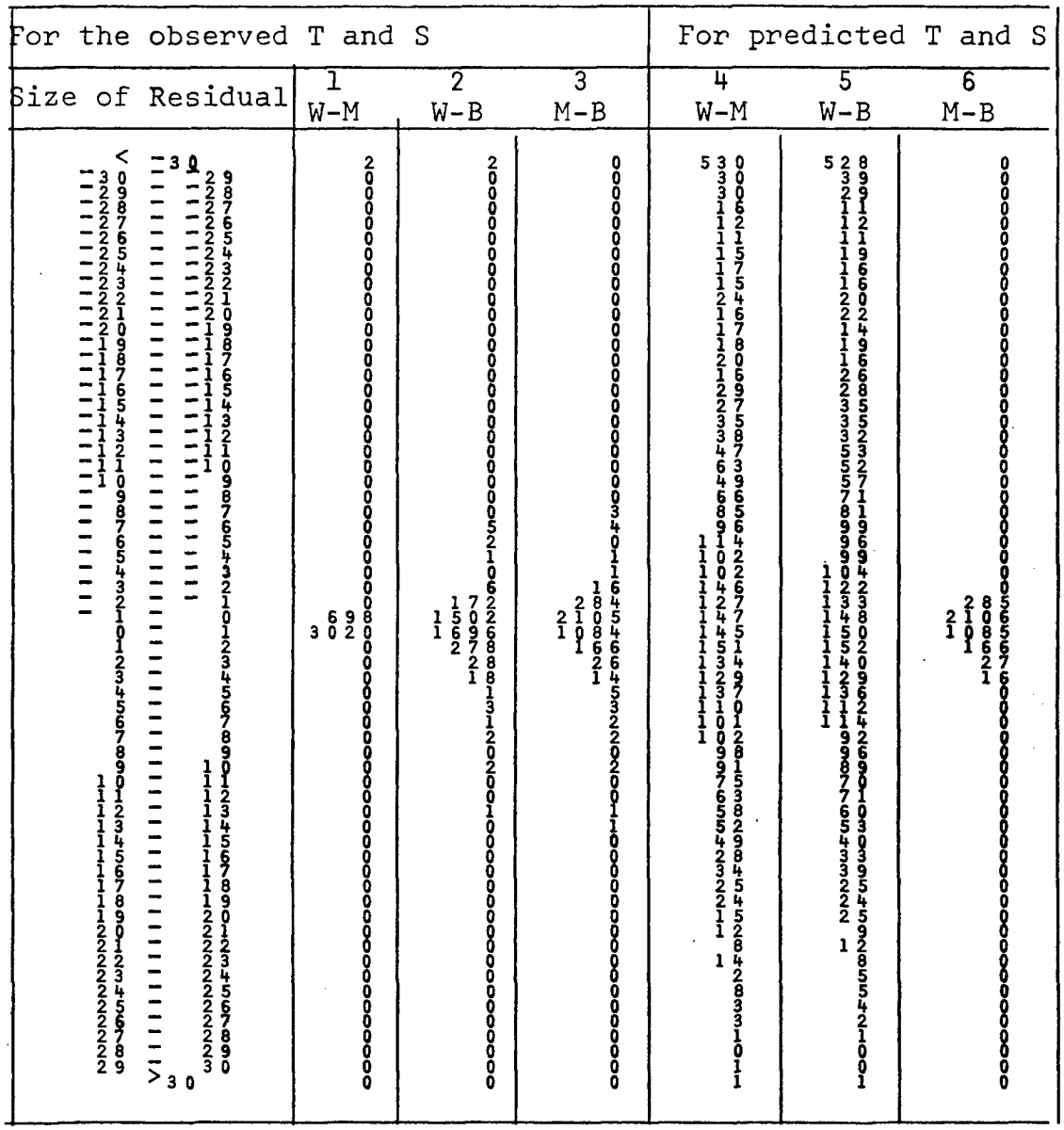


Table II. Residual distribution densities for sound velocity equations.

regression sound velocity equation using predicted temperature and salinity values. If distribution 3 is compared with distribution 6 in table II, it is clear that the regression sound velocity equation predicts very nearly the same as Mackenzie's sound velocity equation.

The nearly normal residual distribution obtained by using the modified Mackenzie equation with the predicted temperature and salinity and the results obtained when these sound velocity values are compared to Wilson's values for the same data, underscores the random error in the data from which the temperature, salinity and sound velocity equations were developed.

Final analysis involved computing a predicted temperature and salinity from their respective regression equations for use in the regression sound velocity equation. The predicted sound velocity from the regression equation ($\hat{SV}$) was compared to Wilson's value for the same data at each observation, forming 3720 residuals (Wilson's sound velocity - regression sound velocity). A plot of these residuals against $\hat{SV}$ for each respective observation revealed a pattern as shown in figure 9. That is, the regression sound velocity equation shows no unaccounted for effect over the range of the dependent variable and indicates a reasonably good fit, as previously noted in the explanation of figure 4. It was observed that 88% of the residuals fell within this horizontal band from +12 m/s to -12 m/s (i.e., no indicated lack of linear or quadratic terms in the sound

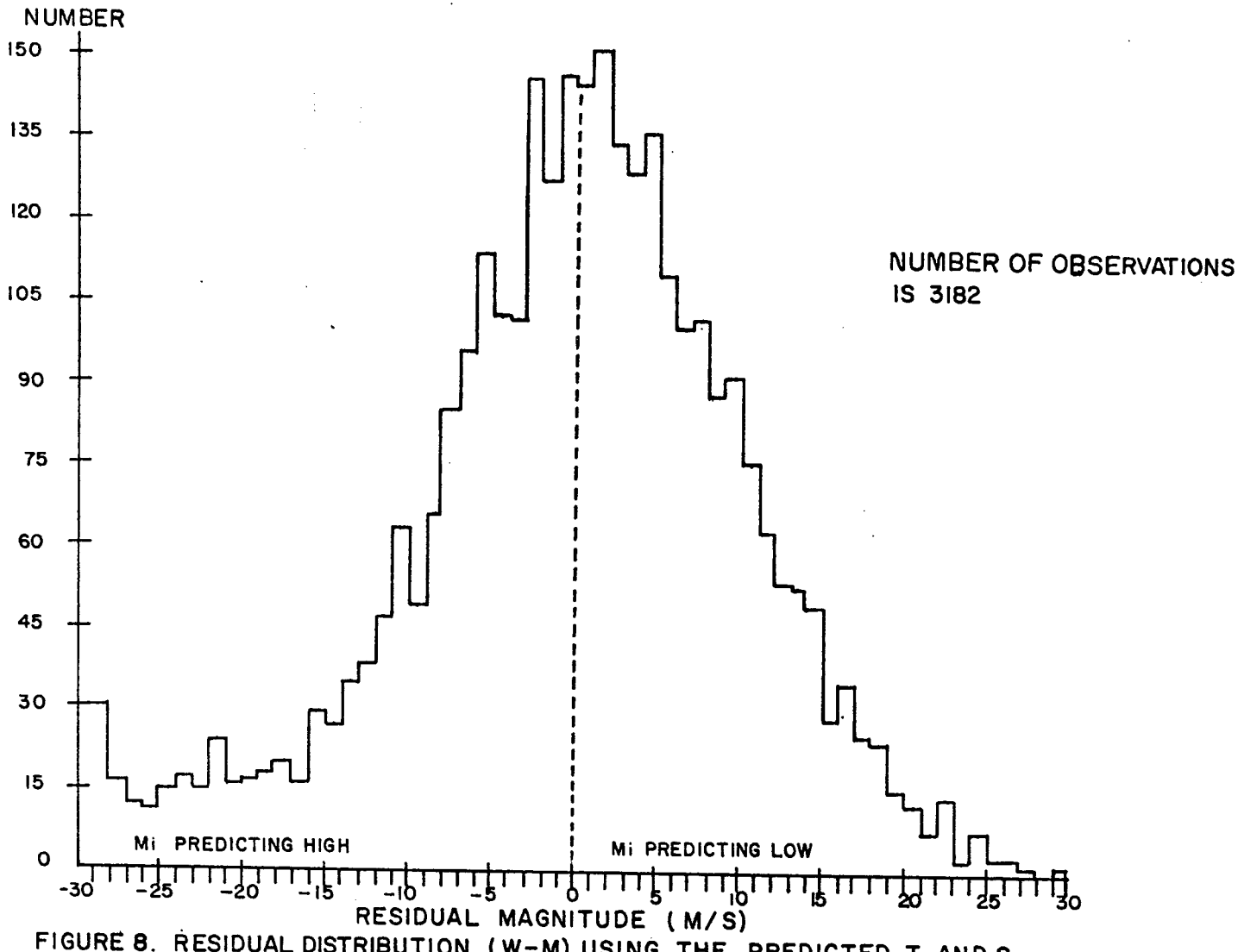


FIGURE 8. RESIDUAL DISTRIBUTION (W-M) USING THE PREDICTED T AND S

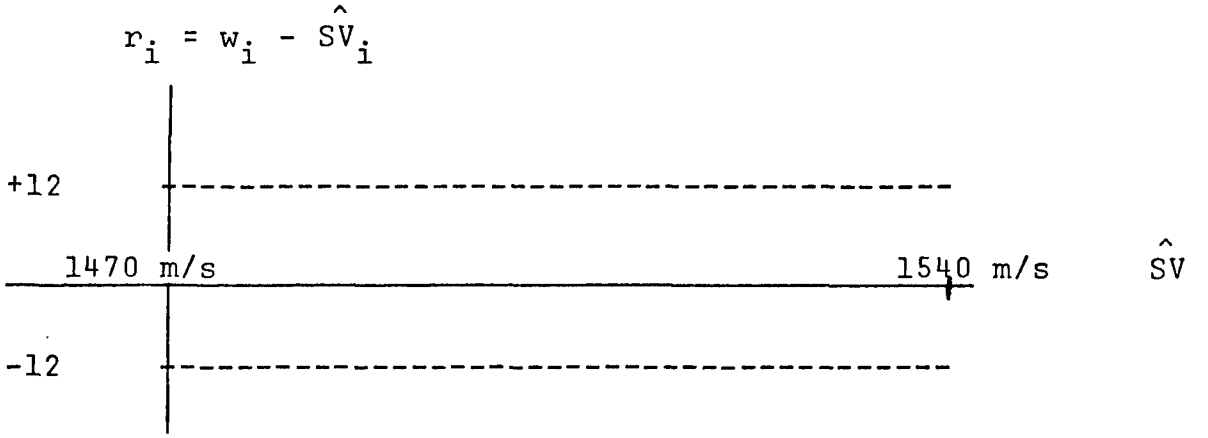


Figure 9. Residual Pattern - plot of residuals against $\hat{SV}$.

velocity model). This residual pattern is what would be expected if the error is random. The analysis presented concerning models 3, 4, and 5 indicates that the error in predictions is random, though large. The prediction of sound velocity without costly instrument measurements of temperature and salinity may require that wider tolerances for error be considered acceptable. For example, based on time and cost saved on instrumental measurements of temperature and salinity, a 90% certainty of being within 5 m/s of the true sound velocity value might be considered adequate.

It is felt that the results of this study are significant enough to warrant application of models 3, 4, and 5 to additional oceanographic data, particularly in squares surrounding the 4^{o} by 4^{o} square used in this investigation.

IV. CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

The problem of determining adequate models for predicting temperature, salinity, and sound velocity has been considered.

Sound velocity values yielded by Wilson's equation described on pages 9 - 12, are considered good enough for use in most scientific work.[2] The Wilson equation, however, is rather complex and requires an excessive amount of calculation. Mackenzie's sound velocity equation, described on pages 6 - 8, is more appealing to use than Wilson's equation because of its simplicity of use. The modification to the reference velocity and depth dependency term, as described on page 18, gives Mackenzie's equation the capability of predicting sound velocities to within ± 1 meter/second of Wilson's equation for all data considered. Distribution 2 of Table II on page 37 shows this result. The Mackenzie equation was therefore concluded to be a convenient and accurate equation from which sound velocity predictions (m_i) could be obtained to compare with the regression sound velocity predictions ($\hat{SV}_i$). Distribution 3 in Table II is formed by considering $m_i - \hat{SV}_i$ for all i, when the observed salinities and temperatures are used in each equation. In contrast, distribution 6 uses the predicted salinities and temperatures in each equation.

Two approaches to the problem of developing prediction equations were used in this investigation. The distinguishing factor between the two approaches is whether depth is included as an independent variable.

Model 1 shown on page 23 and Model 2 shown on page 28 were the primary models considered in the first approach. Depth is not an independent variable in Model 1 or Model 2, therefore a prediction equation for each of the dependent variables temperature, salinity, and sound velocity at each depth plane results.

The results of Models 1 and 2 are discussed on pages 23 and 29 respectively. For each dependent variable, plots of R^2 against depth plane, and σ against depth plane for Model 1 appear on pages 24, 25, and 26. In general, all measures of adequacy as described on page 21, and an examination of residuals (actual - predicted) for each question, fail to substantiate the regression equations yielded by models 1 and 2 as adequate for predictive purposes.

The second approach used in the study was to consider the general situation where depth was included as one of the independent variables. Clearly, this resulted in only one regression equation for each dependent variable temperature, salinity, and sound velocity which represents the data over all depth planes. Data manipulation and analysis of results is much faster if one equation can be found to represent the data over all depth planes, rather than over only one depth plane.

Within the second approach, there were two ways to build the models. First a large model of the form $Y = \Sigma \beta_i X_i + \varepsilon$ could be designed. In using this model, only the dependent variable would be changed. This model would

therefore be used three times. Secondly, three individual models of form

$$Y_T = \sum_{i=1}^{n} a_i X_i + \varepsilon, \quad Y_S = \sum_{i=1}^{n} b_i X_i + \varepsilon, \quad Y_{SV} = \sum_{i=1}^{n} c_i X_i + \varepsilon$$

for temperature, salinity, and sound velocity, respectively, could be developed.

It was concluded in an extensive trial and error model building process, in the search for suitable regression models, that the individual character of the dependent variables required individual models, rather than one large model from which all equations could be derived. The salinity model (model 3), temperature model (model 4), and sound velocity model (model 5) shown on pages 31, 32, and 33, respectively, are the models which gave the best results in the analysis applied.

The salinity equation, obtained from model 3, is a function of latitude, longitude, and depth. The temperature equation, obtained from model 4, is a function of latitude, longitude, depth, salinity, and day-of-year. The final temperature model also included the terms of the model proposed by Anderson[9] for predicting sea surface temperature which also accounts for seasonal variation. The sound velocity equation, obtained from model 5, is a function of latitude, longitude, depth, temperature, and salinity.

When using the prediction equations to arrive at a sound velocity, the following procedure was used.

Salinity may be calculated from values of latitude, longitude, and depth. These are independent variables whose values may be chosen by the user. Once the salinity value is known, and a particular day of year is specified, then a temperature value may be computed. Now both salinity and temperature are defined. These are the only two values that must be known to compute a predicted sound velocity value from either Mackenzie's modified sound velocity equation or the regression sound velocity equation.

For purposes of comparison, the following five sound velocity values were found at each observation of latitude, longitude, depth, temperature, and salinity: Wilson's sound velocity value, Mackenzie's sound velocity and the regression sound velocity using the observed temperature and salinity, and finally Mackenzie's sound velocity and the regression sound velocity using the predicted temperature and salinity.

An assumption that Wilson's sound velocity values were the most accurate, provided a standard of comparison for the sound velocity calculations from Mackenzie's equation and the regression equation. For example, using an observed temperature and salinity, a sound velocity value was calculated from Mackenzie's equation. This sound velocity value was then subtracted from Wilson's value calculated from the same data, and the difference $(w_i - m_i)$ was observed. This was performed at each of the 3720 data points.

Distribution 1 of Table II was formed to see how these residuals were distributed about Wilson's predictions. If

the residual happened to be of magnitude .9 m/sec, the count of all residuals falling in the interval 0 - 1 m/sec was increased by one. Distribution 4 Table II was formed in the same manner using Mackenzie's equation with predicted temperature and salinity. Similar distributions (No. 2 and No. 5 - Table II) were formed regarding the regression sound velocity predictions for observed, as well as predicted temperature and salinity. Two additional distributions (No. 3 and No. 6 - Table II) compare Mackenzie's sound velocity predictions to the regression sound velocity predictions for observed then predicted temperature and salinity, respectively. The six distributions described above are summarized in Table II and reveal some interesting points about the sound velocity equations and their predictive abilities.

When using the observed (instrumental) temperature and salinity in calculating sound velocity from a given equation, Wilson's, Mackenzie's and the regression sound velocity equations all predict sound velocity values very close to one another as distributions 1, 2, and 3 of Table II point out. The regression sound velocity equation resulting from model 5, however, is simpler in form and easier to use than Wilson's equation or Mackenzie's equation.

The residual distributions (No. 4 and No. 5 - Table II), obtained by using predicted temperatures and salinities in computing sound velocity values from Mackenzie's equation and the regression sound velocity equation, are encouraging in that they are nearly normal about Wilson's sound velocity

predictions as shown in figure 8. This form of residual distribution underscores the random error in the data from which the regression equations were developed, and enhances the feasibility of predicting sound velocity without the need for on location, instrument measurement of temperature and salinity.

Figure 9 shows a plot of the residuals $(w_i - \hat{SV}_i)$ against the dependent variable predictions $(\hat{SV}_i)$ for distribution 5, according to the analysis described on pages 21 and 22. Figure 9 differs from figure 8 in that figure 8 is a plot of number of residuals versus magnitude of residual; figure 9 is a plot of magnitude of residual versus magnitude of the dependent variable value $(\hat{SV}_i)$. This plot extends over the entire range of the dependent variable. The plot in figure 9 is that of case A of figure 4, page 22. The residual pattern is roughly a horizontal band, indicating no significant unaccounted for effects (linear or quadratic) in the model over the range of the dependent variable. Since the plot of $(w_i - \hat{SV}_i)$ versus $\hat{SV}_i$, for all i, is a horizontal band, the prediction equation $(\hat{SV})$ is predicting as would be expected if the errors in the raw data for which $\hat{SV}$ was developed, were random.

The regression sound velocity predictions obtained by using predicted salinities and temperature, are not as good as might be desired or needed for use in scientific work. Distribution 5 of Table II shows 528 cases where the regression sound velocity equation predicted values 30 m/sec

previous runs. In addition, the residuals were quite stable. These results substantiated the thought that models 3, 4, and 5 would produce acceptable results if the bad data were removed. Based on these results, it appears feasible that the need for on-location observations of salinity and temperature might be eliminated in the future.

In future work on this topic, some data screening device should be implemented to filter out obvious errors before the final prediction equations, particularly for salinity and temperature, are developed. This would improve the predictive ability of the salinity and temperature equations and thus improve the regression sound velocity predictions.

One such data screening device, which might be used in future investigations, is suggested by Anderson[9]. He proposes that a regression equation be fit to all raw data available as was done in this study. The residuals (observed - predicted) would then be examined. If the residual is ± 2 standard deviations from the mean, that data will be used in further analyses, if not, that data point will be eliminated from further consideration. A regression equation is then fit to the remaining data. This procedure has the facility of immediately identifying erroneous data or gross instrument error.

An alternative to the above data screening procedure would be to compute the mean and standard deviation of the data set in question, then eliminate all data which falls

outside ± 2 or ± 3 standard deviations from the mean. A regression equation could then be fit to the remaining data.

A number of 2° by 2° and 4° by 4° squares adjacent to the area 36° - 40°N latitude and 68° - 72°W longitude were examined. The resulting prediction equations were quite similar in form to those determined for the original square. However, the coefficients of the independent variables were obviously somewhat different. In general, the prediction equations for salinity, temperature and sound velocity in the surrounding areas produced results that were quite good.

For future study on this topic, analysis similar to that discussed in Chapter III of this study, should be performed on several additional $2^{\circ} \times 2^{\circ}$ or $4^{\circ} \times 4^{\circ}$ squares surrounding the area 36° - 40°N latitude and 68° - 72°W longitude. Based on the results from a number of surrounding squares that were examined in this study, the resulting regression equations should be similar to the ones resulting from models 3, 4, and 5 described in Chapter III. These regression equations could then be examined for patterns and possibly generalized equations for salinity, temperature, and sound velocity would become evident which could be applicable to a much expanded oceanographic area.

Physical characteristics of the oceanographic environment are difficult to represent with rigid equations, as is possible in many areas of the physical sciences, because of their dynamic character. The laws of nature, however, are characterized by certain patterns and this environment will eventually be represented too.

BIBLIOGRAPHY

1. MACKENZIE, K.V. Formulas for the computation of sound speed in water. J. Acoust. Soc. Am. 31, 1067(1959).

2. WILSON, W.D. Speed of sound in sea water as a function of temperature, pressure and salinity. J. Acoust. Soc. Am. 32, 641(1960).

3. YERGEN, WALTER E. A rapid access tape format for oceanographic station data. Exploratory Oceanography Division, Report #67-14, March 1967.

4. HECK, N.H. and J. H. SERVICE. Coast and Geodetic Surv. Special publ. No. 108(1924).

5. MATTHEWS, D.J. Hydrographic Dept., British Admiralty, Report HD282, First edition (1927).

6. KUWAHARA, S. J. Astron. and Geophysics 16, 1(1938). Reprinted in Hydrographic Rev. 16, No. 2, 123(1939).

7. DEL GROSSO, V.A. U.S. Naval Research Laboratory Report No. 4002 (1952).

8. YERGEN, WALTER E. Oceanographic Analysis Division Report No. 0-47-64 (1964).

9. ANDERSON, E.R. U.S. Navy Electronics Laboratory Report No. 1427 (1967), Sea-Surface Temperature Estimation.

10. VAN VLIET, C.J. U.S. Navy Electronics Laboratory Report No. 1256 (1965), Sea-Surface Temperature Estimation.

11. ANDERSON, E.R. U.S. Navy Electronics Laboratory Report No. 1429 (1967), Sea-Surface Temperature Estimation.

12. DRAPER, N.D. and H. SMITH. (1967) Applied Regression Analysis. Wiley, New York, p. 88-91.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

U. S. NAVAL OCEANOGRAPHIC OFFICE

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

CRITICAL STUDY OF APPROXIMATING FUNCTIONS (AND METHODS) AS APPLIED TO OCEAN STATION DATA. PROJECT I REGRESSION ANALYSIS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Informal Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Dr. Billy E. Gillett

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1970	51	12

8a. CONTRACT OR GRANT NO.
N62306-68-C00241

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

IR NO. 70-36

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

U. S. Naval Oceanographic Office
Washington, D. C. 20390

13. ABSTRACT

The stepwise multiple regression technique is used in a model building process to develop predictors of temperature, salinity, and sound velocity as functions of geographical location, time, and depth. Models which give reasonable results are obtained through successive trials using higher order terms of the independent variables. The model for sound velocity yields values which are nearly identical to the Wilson sound velocities contained in the ocean station file and values computed using a modified version of the MacKenzie equation.

The distribution of residuals resulting from comparisons of the Wilson equation sound velocities to those obtained from the regression model (both computed from actual temperature and salinities) shows that 98% fall within the range of ±2 m/sec. A comparison of the regression model sound velocity values computed from regression predictions of temperature and salinity with the Wilson values shows that 88% of the residuals fall in the range of ±12 m/sec.

The results, which are valid for the 4° square centered at 37.5° North latitude and 69.5° West longitude, are discussed in terms of the statistical significance of the distribution of the residuals. Since the physical characteristics of the area selected are rather complex, the application of this technique to other parts of the ocean is recommended.

This work was performed under NAVOCEANO Contract No. N62306-68-C00241 by Dr. Billy E. Gillett, Department of Statistics and Applied Mathematics, University of Missouri in Rolla, Missouri.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
CRITICAL STUDY OF APPROXIMATING FUNCTIONS (AND METHODS) AS APPLIED TO OCEAN STATION DATA. PROJECT I REGRESSION ANALYSIS						