Tellus (1977), *29*, 545–560

Objective analysis and classification of oceanographic data

By JOHN B. JALICKEE, *Center for Experiment Design and Data Analysis, Environmental Data Service, National Oceanic and Atmospheric Administration, Washington D.C. 20235, U.S.A.*

and DOUGLAS R. HAMILTON, *National Oceanographic Data Center, Environmental Data Service, National Oceanic and Atmospheric Administration, Washington, D.C. 20235, U.S.A.*

(Manuscript received November 10, 1976)

ABSTRACT

A new approach to the analysis and classification of oceanographic data is presented. The technique is an empirical one, based on the singular decomposition theorem, for characterizing temperature–salinity–depth profiles as in water-mass analysis. Complementary to the profiles are station or cast-dependent coefficients, which show similarities and differences according to the space–time distributions of the individual stations in the data set. These coefficients are used for classifying the individual stations into groups having similar profiles. Results of applying the new method to a set of data from ocean stations off the coast of Oregon are given.

1. Introduction

Studies related to the oceans often require a precise, objective description or characterization of the physical nature of the region of interest. Such a characterization is usually in the form of expected values and variations of one or more parameters within several geographic areas, time periods, and depths. Statistics describing a data population are most accurate and useful if the population is homogeneous, but it is difficult to identify homogeneous populations in the dynamic ocean. One approach often used is to choose geographic subareas and seasonal time periods in such a way that each exhibits a unique set of attributes. Inherent in this approach, however, is the assumption that oceanographic features conform to static geographic and time boundaries.

Another possible approach is to group data according to similarity of various features, such as the shape of temperature–salinity–depth (T–S–Z) curves, as in water-mass analysis. If this can be done, the variability within each group is minimized, and the expected values of the parameters in each group are more reliable than the expected values of the total population. Also, if the grouping succeeds in characterizing oceanographic regions, natural boundaries or transition zones in time and space will be apparent in time and location plots of the stations of each group.

Described here is a numerical technique, called Asymptotic Singular Decomposition (ASD), which has been developed at the National Oceanic and Atmospheric Administration for the analysis of complex data sets. The singular decomposition theorem is discussed in Section 2, and an empirical model based on the theorem, specifically tailored for analysis of T–S–Z data, is developed in Section 3. A typical data set for analysis from ocean stations off the coast of Oregon is described in Section 4, the diagnostic value of the method is demonstrated in Section 5. Other uses of the analysis scheme are discussed in Section 6. In Section 7 the objective classification procedure is derived and results from the application of this method to the oceanographic data set are presented in Section 8, followed by a summary and conclusions in Section 9.

2. Singular decomposition

The basic mathematical concept on which the analysis is based is the singular decomposition of a matrix. Let $\mathbb{A} = \{A_{ij}\}$ be an $N \times M$ matrix $N \leq M$, then the singular decomposition of $\mathbb{A}$ is

$$A_{ij} = \lambda_1 f_i^{(1)} g_j^{(1)} + \lambda_2 f_i^{(2)} g_j^{(2)} + \cdots + \lambda_N f_i^{(N)} g_j^{(N)} \quad (1)$$

The numbers $\lambda_1, \lambda_2, \ldots, \lambda_N$ are the singular values of $\mathbb{A}$, with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0 \tag{2}$$

Elements $f_1^{(l)}, f_2^{(l)}, \ldots, f_N^{(l)}$ and $g_1^{(l)}, g_2^{(l)}, \ldots, g_M^{(l)}$ form the lth left and right singular vectors of $\mathbb{A}$ respectively. These vectors are mutually orthonormal:

$$\sum_{i=1}^{N} f_i^{(l)} f_i^{(m)} = \sum_{j=1}^{M} g_j^{(l)} g_j^{(m)} = \begin{cases} 0 & l \neq m \\ 1 & l = m \end{cases} \tag{3}$$

The value of the singular decomposition in this study lies in the fact that the first k terms in the expansion form an optimal approximation to the original matrix in the sense of the least squares. Let $\hat{A}_{ij}^{(k)}$ denote an approximation comprised of the first k terms:

$$\hat{A}_{ij}^{(k)} = \sum_{l=1}^{k} \lambda_l f_i^{(l)} g_j^{(l)} \tag{4}$$

Then

$$\sum_{i,j} (A_{ij} - \hat{A}_{ij}^{(k)})^2 = \text{Minimum} \tag{5}$$

with respect to all possible choices of the elements of the terms in expansion of $\hat{A}_{ij}^{(k)}$. These terms are also ordered in terms of the variance each explains.

$$\left(\lambda_l^2 \Big/ \sum_{k=1}^{N} \lambda_k^2\right) * 100\%$$

of the total variance of the entire array is explained by the lth term and since the λ's are ordered, so are the individual variances explained by them.

This scheme of successive approximations can be shown to be equivalent to principal component analysis (Whittle, 1952), but differs in two respects: (i) the conjugate pairs of vectors are on an equal footing and both can be given physical interpretation, and (ii) the computational methods for calculating the terms in the singular expansion are considerably more accurate and reliable than those

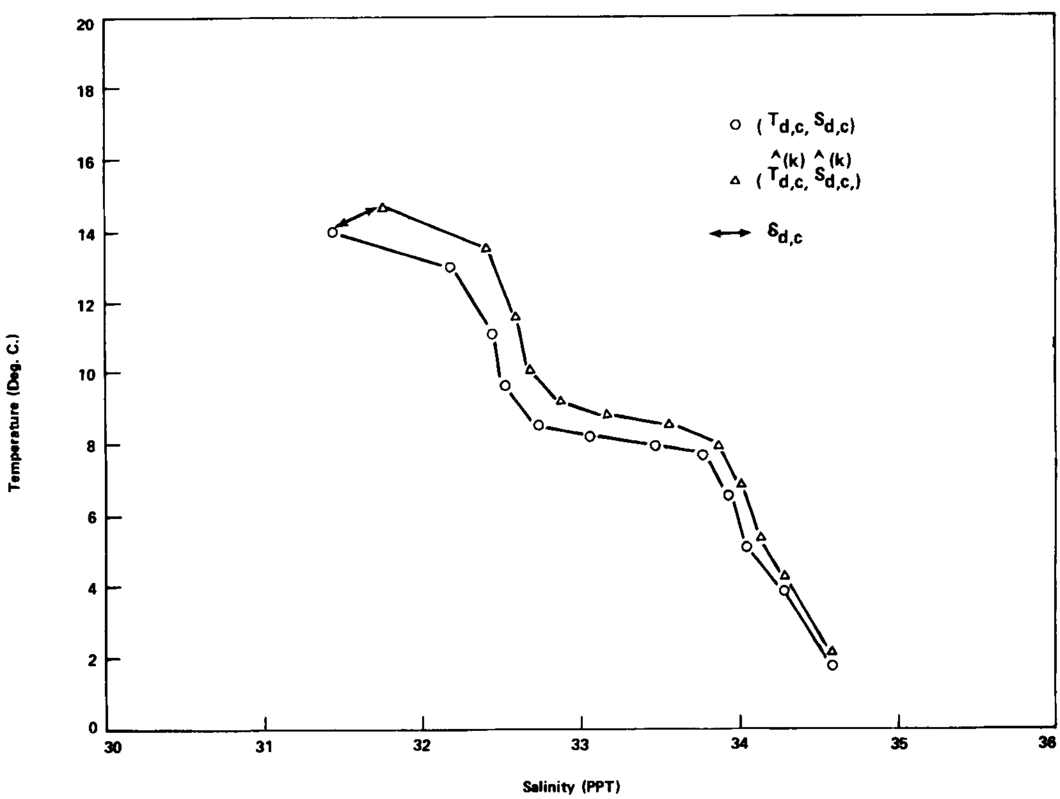


Fig. 1. Example of the minimization function described in eq. (10).

commonly used in principal component analysis. Good (1969) has made these points and described other uses of singular decomposition. Golub & Reinsch (1971) have pointed out the pitfalls of calculating the terms of the expansion by "equivalent" methods and have developed an efficient and accurate method for calculating the terms directly. The Asymptotic Singular Decomposition (ASD) method, however, which has been described in detail by Jalickee (1976), is especially designed for calculating the leading terms of the expansions of very large matrices.

3. Model of ocean data

For the analysis of T–S–Z data the basic definition of the data matrix was modified as follows: Let $T_{d,c}$ and $S_{d,c}$ denote the temperature and salinity respectively at depth level d, $d = 1, 2, \ldots, N$ for the cth station, $c = 1, 2, \ldots, M$. Then the T–S–Z matrix $\mathbb{A}$ is defined by:

$$\begin{aligned} A_{1,d,c} &= T_{d,c} \\ A_{2,d,c} &= S_{d,c} \end{aligned} \tag{6}$$

and the singular decomposition is written in the form:

$$A_{i,d,c} = \sum_{l=1}^{2N} \lambda_l f_{i,d}^{(l)} g_c^{(l)} \tag{7}$$

This matrix is then considered as a $2N \times M$ matrix and is decomposed by an appropriate computational procedure.

Now, if we define

$$T_d^{(l)} = \lambda_l f_{1,d}^{(l)} \tag{8a}$$

and

$$S_d^{(l)} = \lambda_l f_{2,d}^{(l)} \tag{8b}$$

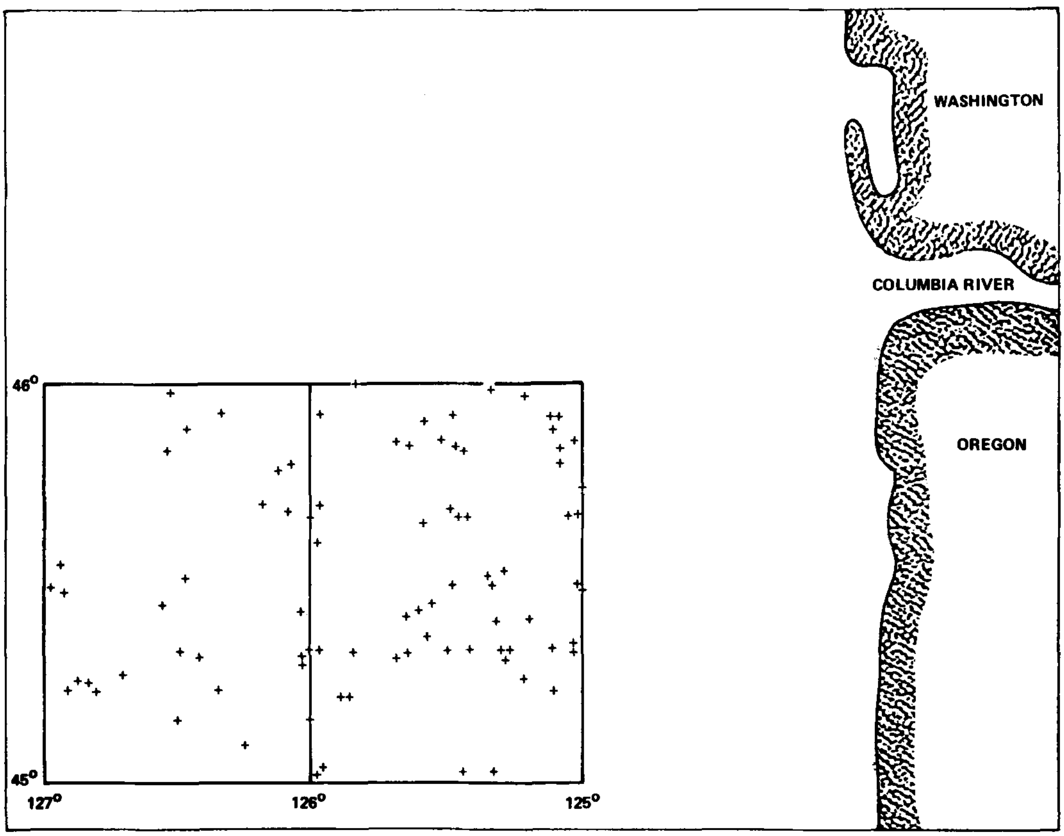
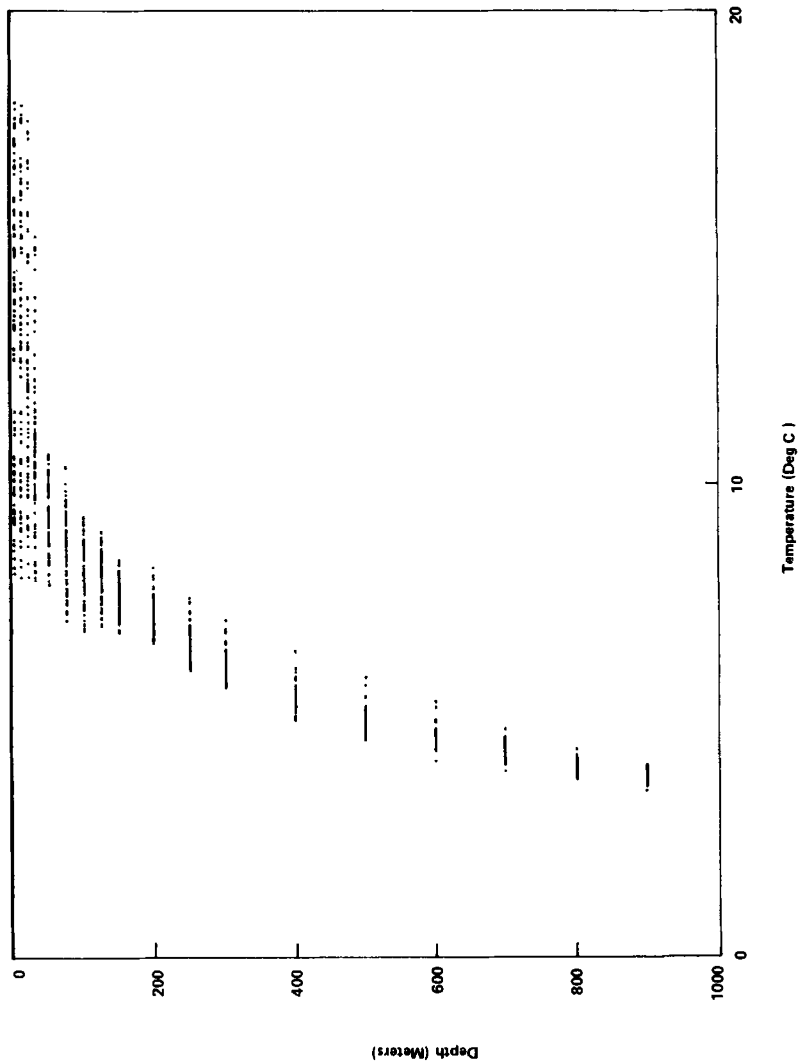


Fig. 2. Locations of the station data used in the analysis.



a

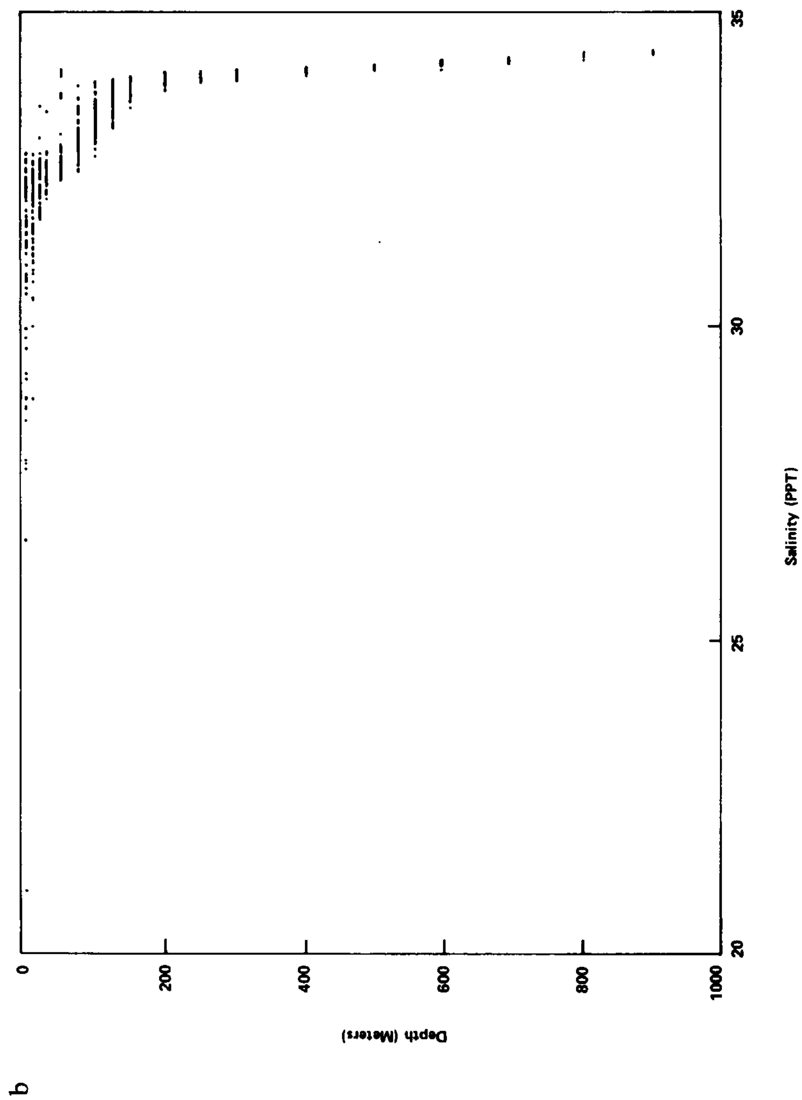


Fig. 3. Temperature–depth (a) and salinity–depth (b) distributions.

we can rewrite eq. (7) in a more descriptive form:

$$(T_{d,c}, S_{d,c}) = \sum_{l=1}^{2N} (T_d^{(l)}, S_d^{(l)})\, g_c^{(l)} \tag{9}$$

where the new left singular suggests characteristic T–S–Z profiles. The g_c's will be called station coefficients.

To show how these characteristic functions are optimum we denote by $(\hat{T}_{d,c}^{(k)}, \hat{S}_{d,c}^{(k)})$, the kth approximation, the sum of the first k terms of eq. (9). Then the approximation satisfies:

$$\sum_{d,c} \delta_{d,c}^2 = \text{Minimum}$$

where

$$\delta_{d,c}^2 = (T_{d,c} - \hat{T}_{d,c}^{(k)})^2 + (S_{d,c} - \hat{S}_{d,c}^{(k)})^2 \tag{10}$$

The minimum is over all possible choices of the characteristic profiles and station coefficients. This is shown graphically in Fig. 1.

Because the elements in the model are determined empirically, they contain physically significant information related to ocean conditions as embodied in the temperature–salinity data. Also, the terms in the expansion are orthogonal to one another; hence each term represents independent information. Since the terms are ordered according to the variance explained, the higher order terms generally describe finer details of the data set.

The first characteristic profile, $(T_d^{(1)}, S_d^{(1)})$, effectively constitutes the typical profile, since it alone best describes the entire collection of original profiles. Subsequent terms modify the collection of previous terms in describing next-most-important variations. The station coefficients in the expansion modify these profile values to produce a unique approximation to each of the stations. Therefore, a unique set of coefficients exists for each different station. Because the characteristic profiles are the same for all stations, the unique T–S–Z characteristics of a particular station are reflected in the

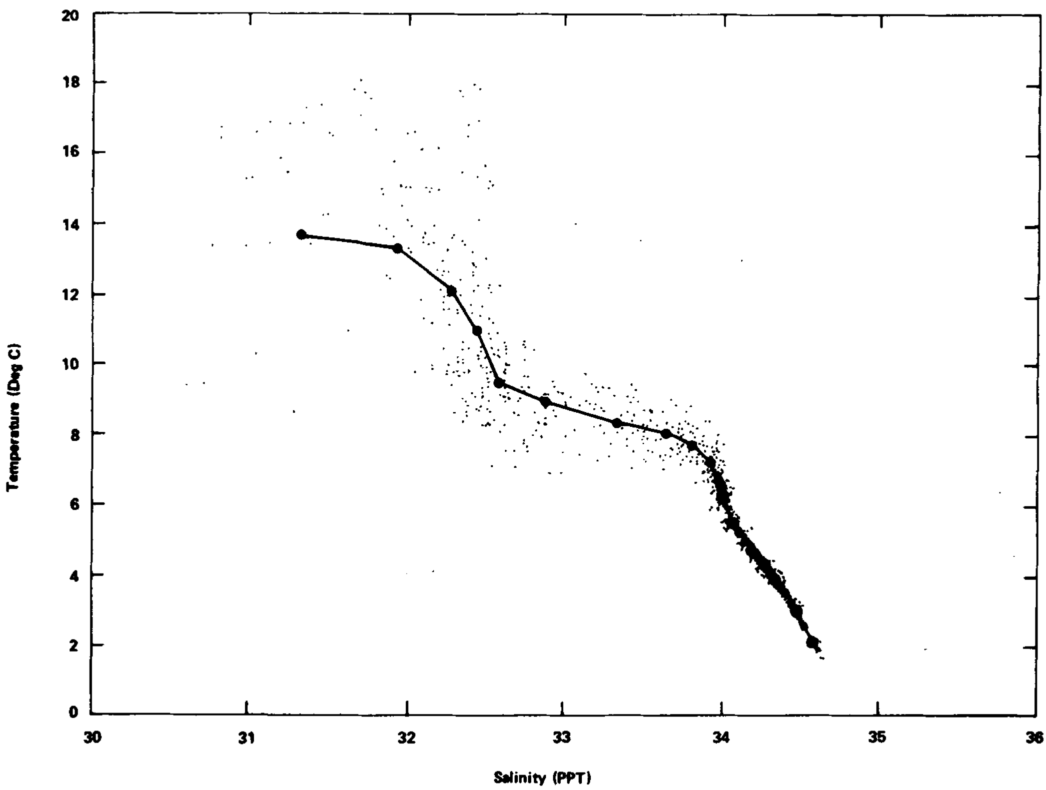


Fig. 4. Temperature–salinity data with the first characteristic T–S– Z profile superimposed.

coefficients of that station. The sets of coefficients for stations with similar T–S–Z profiles will be similar, while the sets corresponding to different profiles will be diverse. This property of the coefficients make them useful as indicators of similarities and differences between casts and constitutes the basis for the objective classification scheme described in Section 7.

4. Data used in the analysis

Data obtained from a group of stations located off the coast of Oregon were selected from the ocean station data archive at the National Oceanographic Data Center to examine characteristics of the ASD technique using oceanographic data. One requirement to applying this technique is that the data must be in matrix form, with all cells filled. For this analysis, the matrix was made up of temperature and salinity values at standard depth levels. Thus, depths are not directly a part of the analysis, but the depth of each datum is known indirectly by its position in the array.

To obtain a sufficient number of stations with enough depth for a meaningful analysis, stations with temperature and salinity values at each standard depth between the surface and 900 m were accepted. This yielded a total of 95 stations in the region 45° to 46° north latitude and 125° to 127° west longitude (Fig. 2). The matrix constructed from these stations consists of 36 rows (18 depths for both temperature and salinity) and 95 columns (one for each station).

The temperature–salinity–depth distributions of the data used in the analysis are shown in Figs. 3 and 4. These stations exhibit characteristics typical of the coastal subarctic Pacific Ocean as described by Dodimead et al. (1963). The problem now is to identify groups within this data set, each of which has a unique temperature–salinity relationship. Before discussing the resulting of the ASD classification, however, we will examine the information obtained from the model functions and coefficients.

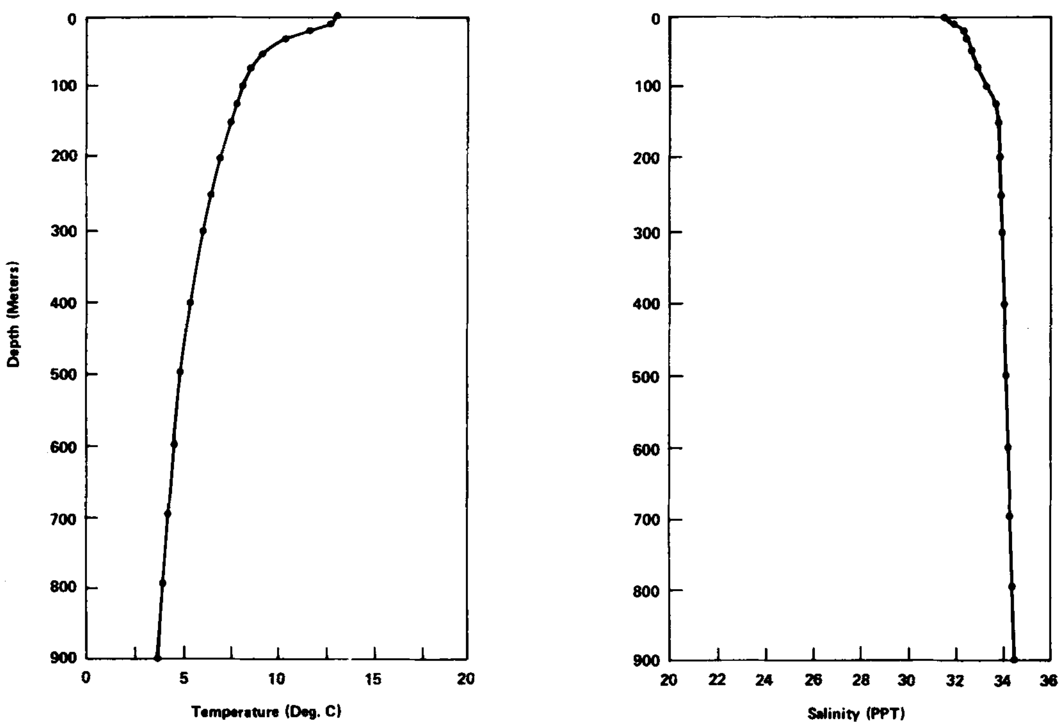
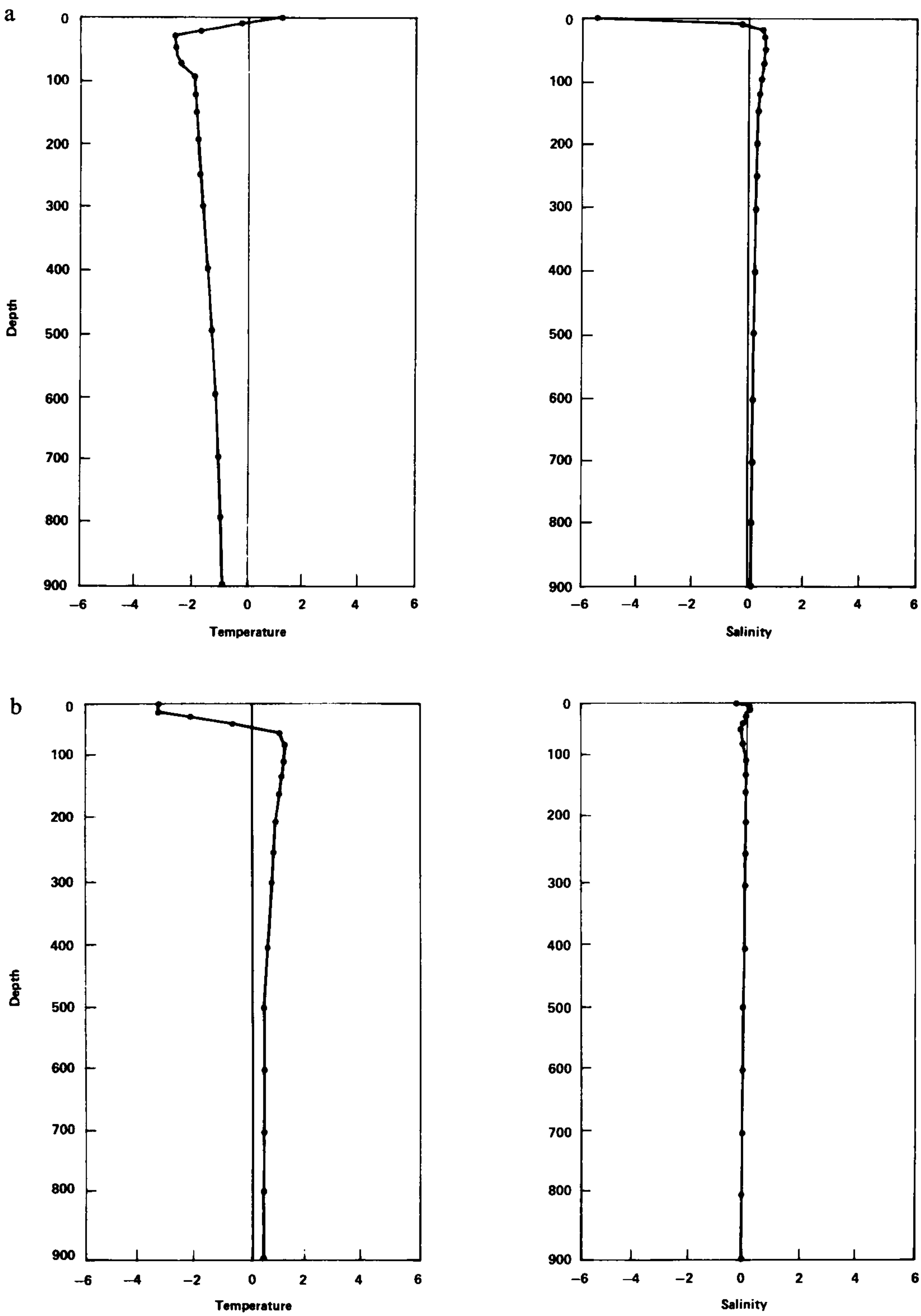


Fig. 5. First characteristic profiles representing the typical depth dependence of temperature and salinity.


a
b
Depth
Temperature
Salinity
0
100
200
300
400
500
600
700
800
900
−6
−4
−2
0
2
4
6

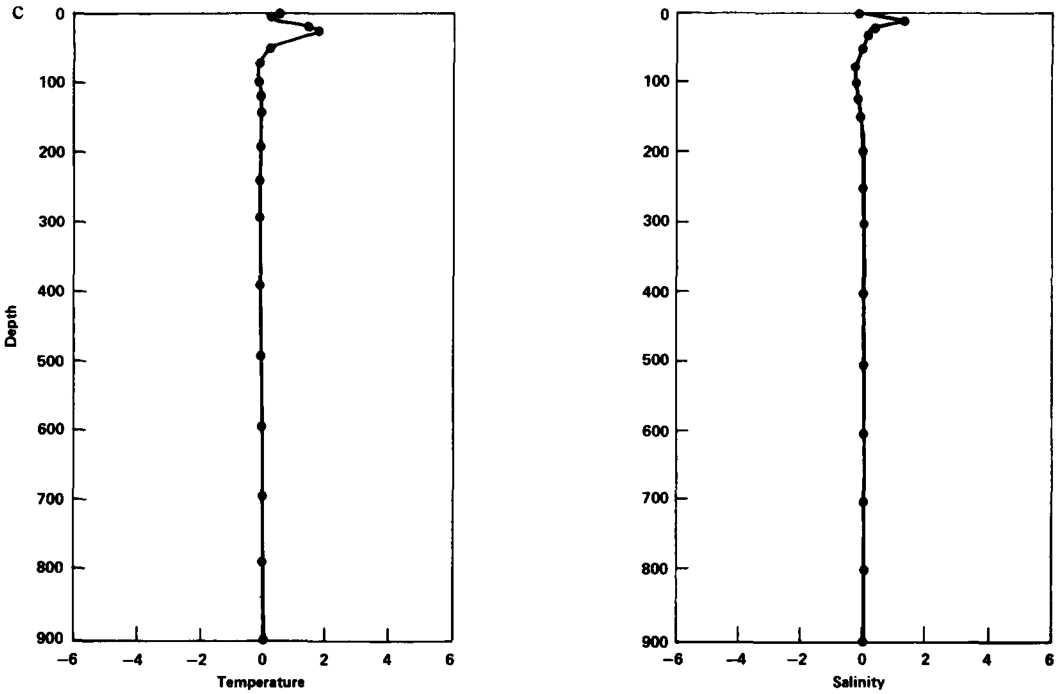


Fig. 6. Characteristic profiles two (a), three (b), and four (c).

5. Physical information obtained from model elements

5.1. *Functions*

Features of the vertical distributions of temperature and salinity of the 95 stations are shown in the profiles computed by ASD. The first profile $(T_d^{(1)}, S_d^{(1)})$, when multiplied by the average station-coefficient $\bar{g}_c^{(1)}$, representative of all stations, provides an estimate of expected average values of temperature and salinity at each depth level, as shown in Fig. 5. These values are also shown in Fig. 4, superimposed on the data used in the analysis. Succeeding functions (Fig. 6) give information about the most significant variations of actual profiles from the first estimate. Each of these, when multiplied by a station coefficient and added to the first function, makes a slight change in the approximation to produce an estimate of a particular station. Estimates of the profiles comprising the data set become more accurate with the addition of each successive term.

Of greatest interest are the second and third functions. The second function contributes a large negative salinity at the surface and fairly large negative temperature values between 20 and 900 m. The third function, on the other hand, primarily modifies temperature values, especially in the upper 20 m. The fourth function contributes a small correction to both salinity and temperature in the uppermost layers. The functions act together in an infinite number of possible combinations to add to or subtract from each other at each depth level.

5.2. *Coefficients*

Coefficients of these functions, computed for this analysis, reveal the similarities and disparities of stations. The first coefficient, when applied to the first function, determines the magnitude of temperature and salinity values in the first approximation, and can therefore be used as an index to the relative magnitudes of individual stations. Similarly, subsequent coefficients determine the magnitude and sign of modifications to the first approximation. The second coefficient of each station is an indicator of the presence or absence of low surface salinity; a station with normal surface salinity (about 32.5‰) would have a second coefficient near zero, but the coefficient of a station

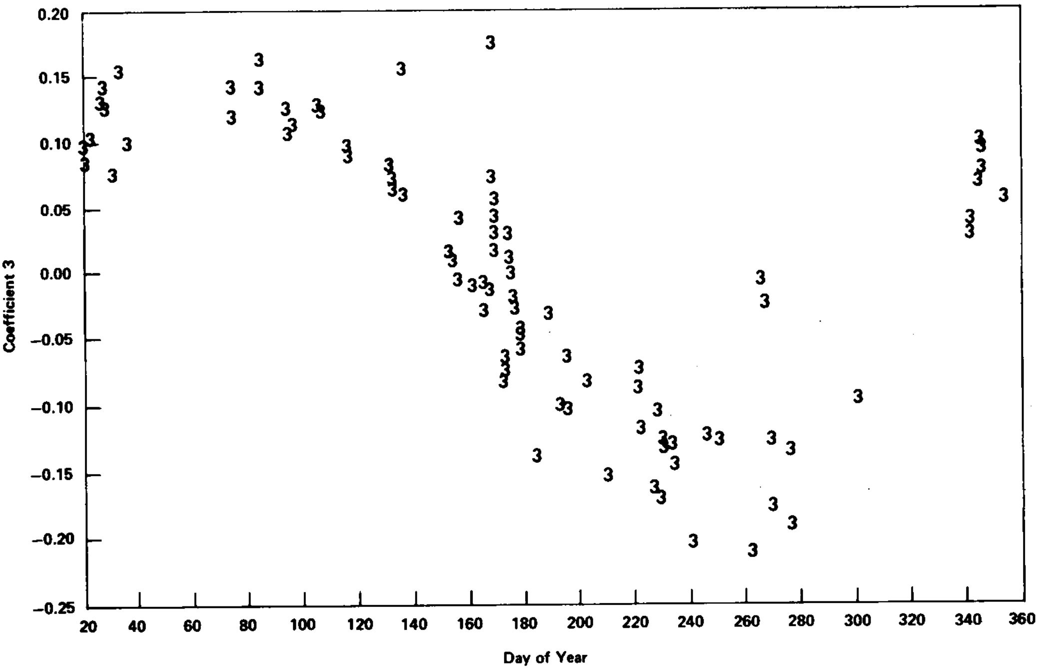


Fig. 7. Day-of-year dependence of station coefficient 3.

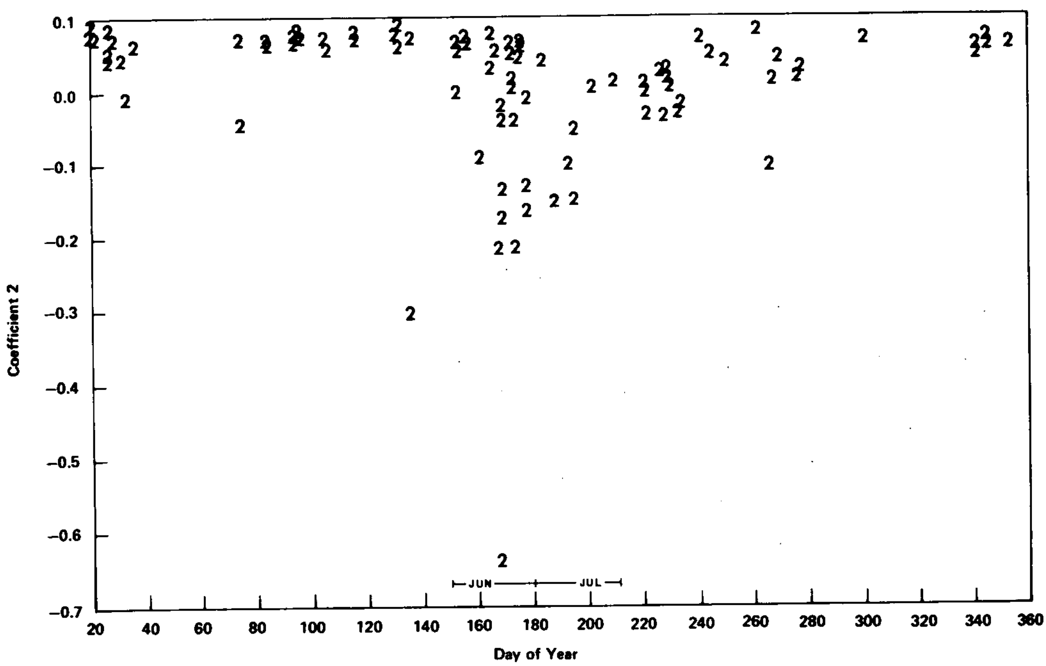


Fig. 8. Day-of-year dependence of station coefficient 2.

with extremely low surface salinity will have a large negative value.

Turning now to the distribution of individual coefficients of all stations as a function of the Julian Day of year of each station, we find that the sine curve pattern of the third coefficients (Fig. 7) is reasonable, since changes made by the third function are mostly to temperature in the upper 20 to 30 m. Obviously, this distribution reflects seasonal temperature change in the surface layer. On the other hand, the distribution of the second coefficient, shown in Fig. 8, is uniform during most of the year, with the exception of a rather sharp peak in June and July. Recalling that the second function modifies surface salinity values and that this region is near the Columbia River, we can assume that the large negative coefficients in June belong to stations taken in the Columbia River plume. In fact, the 15-year (1953–1967) mean of river outflow at Astoria, Oregon, indicates that the normal discharge volume of the river is about 7500 m^3/s during most of the year, but in mid-June there is a peak of about 17,500 m^3/s. There is also a minor peak during February, due to coastal river run-off into the Columbia River, but the plume is generally turned northward along the coast by southerly winds during winter. In summer, however, the plume flows southwestward through the region being studied because of the predominant northerly winds (Barnes et al. 1972).

In addition to time distributions of coefficients, geographic distributions can be useful in the analysis. For example, a plot of the second coefficients during June and July (Fig. 9) shows a correlation of large negative values with the northeast portion. Translating this into salinity, we would not expect to find low surface salinities in regions with positive second coefficients because the second function would add rather than subtract salinity at the surface.

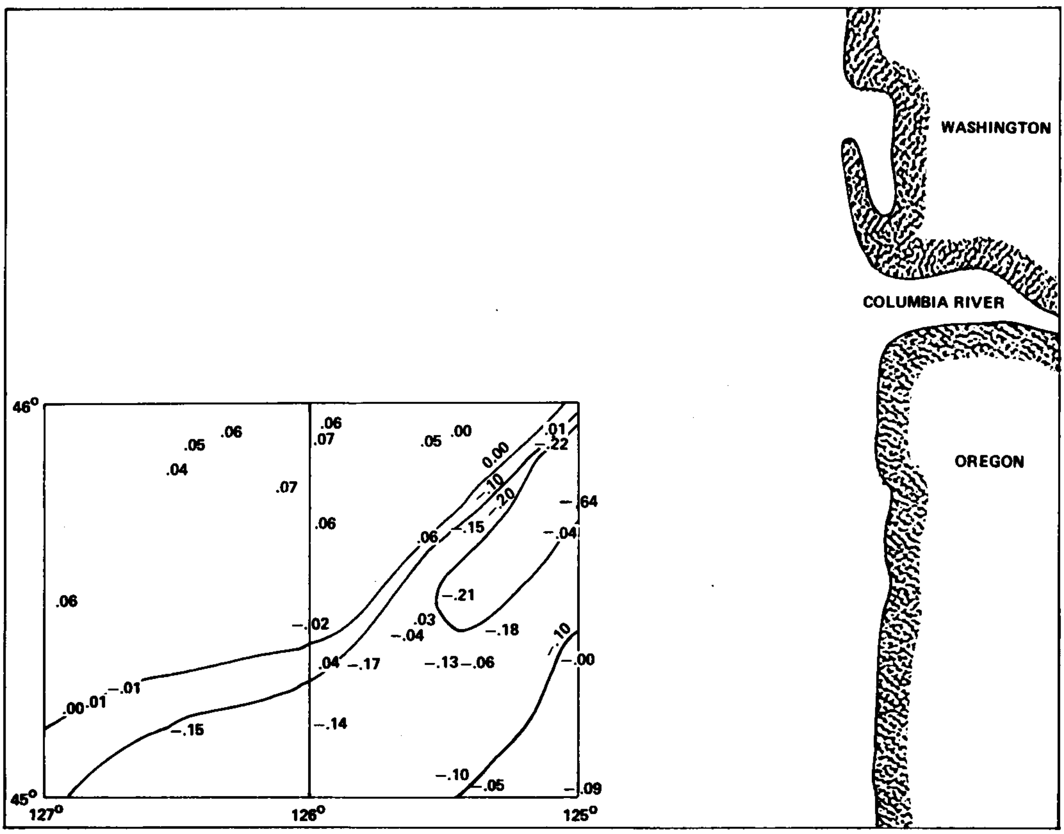


Fig. 9. Contour map of station coefficients for June and July.

6. Other uses of model elements

The model elements are potentially useful for solving a number of other types of problems. In the approximation of data from the model elements, for example, minor irregularities of individual T–S profiles are eliminated. Therefore, smooth data fields can be developed from reconstructed data.

Distributions of coefficients in time and space provide the possibility of data validation, and interpolation on time and location fields. If a station shows values that exceed the normal range of variation, the coefficients of that station will obviously be different from coefficients of normal stations, particularly on plots of coefficient distributions. If there is a period of time or a geographic area for which no data exist, it is possible to interpolate coefficients from existing distributions and create an expected T–S–Z profile. Of course, the reliability of that profile will depend on the continuity of actual ocean conditions and the amount of data surrounding the region of interpolation.

All T–S–Z profiles in the original data set can, of course, be approximated by the model. Therefore, the model can be used as a compressed version of the original data. In this analysis, there were 18 temperature and salinity values per station and 95 stations, or a total of 3420 data values. The model consists of four singular terms, four functions, each having 36 values, and four coefficients for each station, or a total of 528 values. This represents an almost six-fold reduction, and reduction would be even greater for larger data sets.

7. Classification of T–S–Z profiles

In order to determine homogeneous groups within the data set the model elements are rearranged to form a different expansion, consisting of new functions and station coefficients that are particular linear combinations of the original ones. These new functions and coefficients not only approximate each T–S profile, but also represent each of the homogeneous groups. In the original model, the first function was a first-order estimate of temperature–depth and salinity–depth profile shapes for all stations in general. In the new model, however, each function represents the temperature–depth and salinity–depth profiles of one of the groups, the second function being a first-order estimate of T–S–Z in group 2, and so forth. The station coefficients of the new model indicate the degree of association of each station with each of the groups. The T–S–Z profile of a station is most similar to the group whose coefficient is largest in magnitude for that station. Thus, the new model functions provide representative T–S–Z profiles for each group, and the new coefficients indicate the group to which each station belongs.

The expansion, eq. (9), rewritten in terms of the new components, is

$$(\tilde{T}_{d,c}^{(k)}, \tilde{S}_{d,c}^{(k)}) = \sum_{l=1}^{k} (\tilde{T}_d^{(l)}, \tilde{S}_d^{(l)})\tilde{g}_c^{(l)} \tag{11}$$

Here the new station coefficients, $\tilde{g}_c^{(l)}$, are defined in terms of the old coefficients through the transformation defined by the elements $\{\alpha_{l,m}\}$:

$$\tilde{g}_c^{(l)} = \sum_{m=1}^{k} \alpha_{l,m} g_c^{(m)} \tag{12}$$

We require these new coefficients to retain the normalization of the old, so that

$$\sum_{c=1}^{M} \tilde{g}_c^{(l)2} = 1 \tag{13}$$

which implies

$$\sum_{m=1}^{k} \alpha_{l,m}^2 = 1 \tag{14}$$

The new T–S–Z profiles are derived from the old using the inverse transformation described by the elements $\{\beta_{lm}\}$. That is, with

$$\sum_{l=1}^{k} \beta_{l,n}\alpha_{l,m} = \begin{cases} 0 & n \neq m \\ 1 & n = m \end{cases} \tag{15}$$

we have

$$(\tilde{T}_d^{(l)}, \tilde{S}_d^{(l)}) = \sum_{n=1}^{k} (T_d^{(n)}, S_d^{(n)})\, \beta_{l,n} \tag{16}$$

Using the definition of the new factors and the transformations, one can show that eqs. (9) and (11) are in fact identical. We have merely rearranged the terms and not changed the actual approximation.

Our choice of transformation elements is made under the requirement that the new station coefficients be as mutually exclusive as possible. To this end, consider the product

$$(\tilde{g}_c^{(l)} * \tilde{g}_c^{(m)})^2, \quad l \neq m$$

We see that this number is non-negative and identically zero when the groups are exactly mutually exclusive. If one and only one station coefficient, $g_c^{(3)}$ say, is different from zero for a given station, that station belongs to one and only one new characteristic profile or group; in this case, the third group. This is not generally the case, but we attempt to approach this condition by minimizing the sum of all such possibilities through

$$\sum_c \left\{ \sum_{l} \sum_{\substack{m \\ l \neq m}} (g_c^{(l)} * g_c^{(m)})^2 \right\} = \text{Minimum} \qquad (17)$$

with respect to the set of α's under the conditions of normality in eq. (14). This method is quite similar to methods of factor rotations in factor analysis. The procedure developed by Powell (1964) was used to achieve the nimimization.

Transformation of the original model produced four homogeneous groups, which are discussed below.

8. Results

The ideal classification would be one in which the group of every station is clearly identifiable. In reality, however, some stations will have characteristics that overlap between two or more groups. Therefore, each group may include some stations that do not seem to belong, but each is, in general, an entity with unique characteristics. The temperature–salinity traits of the first group (Fig. 10a) are typical of winter coastal subarctic waters, with a relatively cool, low salinity (32.0–32.5‰) mixed layer about 50 m deep. Stations in this group are found in all parts of the region from December through the early part of June.

Stations in the second group (Fig. 10b) are typical of summer coastal subarctic waters. The salinity–depth distribution is similar to that of the first group, but the temperature–depth distribution shows higher values in the surface mixed layer. Stations with these characteristics are distributed throughout the area from June through October (and perhaps into November, although no station data are available for that month). The first two groups represent normal seasonal T–S conditions in this region; there is no geographic correlation of any consequence in these groups, and together they span all months of the year. Groups 3 and 4, however, reflect departures from the norm due to the influx of Columbia River water.

Salinity values of stations in the third group (Fig. 10c) are extremely low at the surface (less than 30‰) and nearly normal at 10 m, which indicates the presence of a thin layer of relatively fresh water. Stations in this group cover almost exclusively the month of June and are located in a fairly well-defined northeast-to-southwest band.

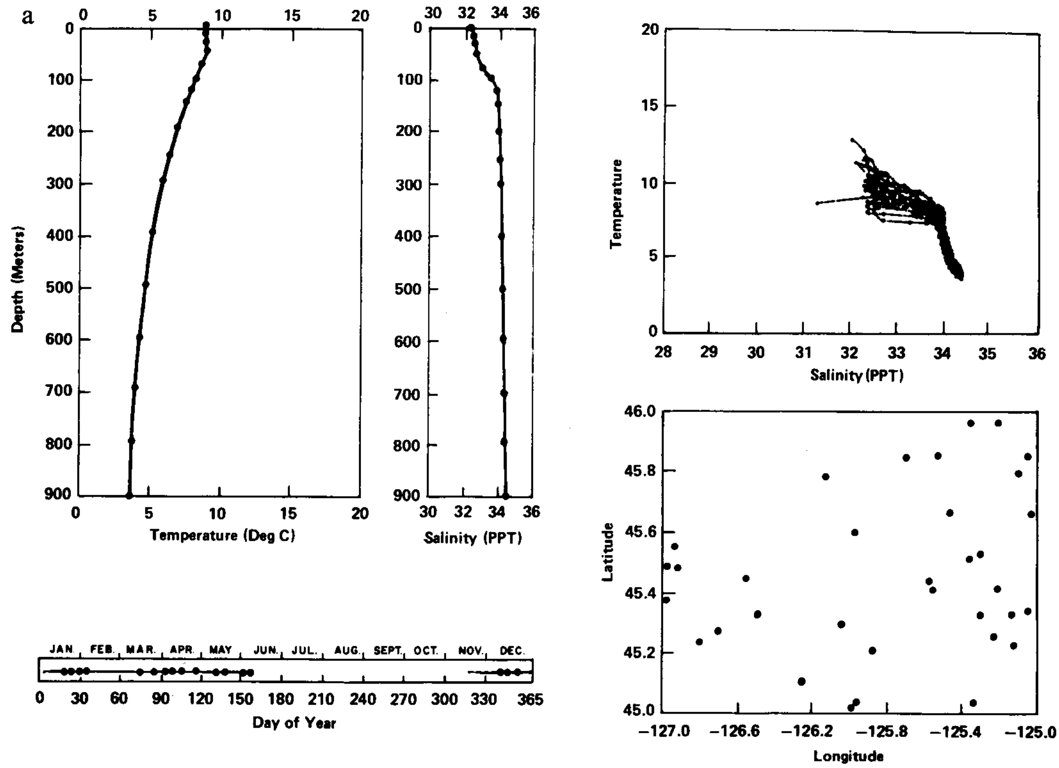
Temperature–salinity–depth characteristics of stations in the fourth group (Fig. 10d) indicate the presence of modified Columbia River outflow. Both temperature and salinity profiles show a mixed layer of 10 m depth in which the water is warm and of moderately low salinity (about 30‰). At 20 m depth, however, salinity values are generally higher than 32‰. Although the stations in group 4 are geographically more widespread than those in group 3, they tend to be located primarily in the eastern part of the region. The distribution of these stations in time is also more widespread, covering the months June through September. The presence of one station in March is perhaps the result of the winter peak in Columbia River outflow.

Examination of the T–S plots of the four groups in Fig. 10 shows that the stations have been successfully classified into homogeneous groups by the ASD method. Furthermore, as a result of the classification, the region of interest can be definitively described in terms of the T–S–Z characteristics and the geographic and time distributions of stations in each of the four groups.

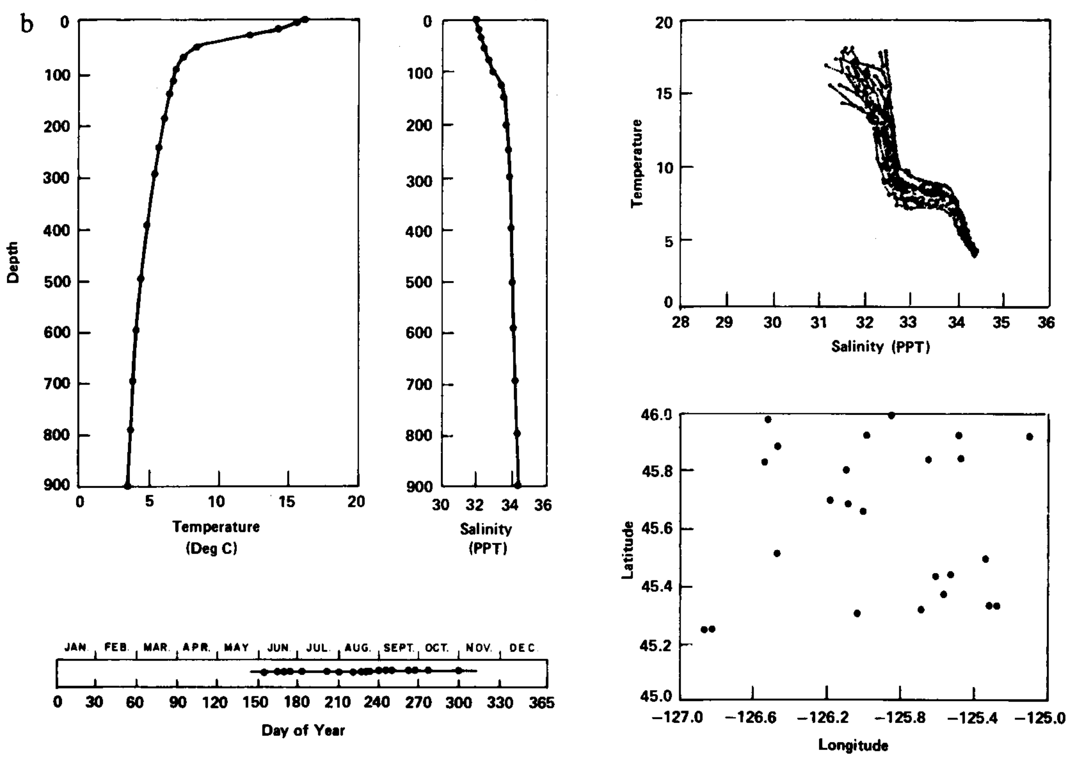
9. Summary and conclusions

We have described and illustrated an approach to the analysis and classification of oceanographic data. Using the technique of singular decomposition, important features of the data are extracted and represented in the form of optimum characteristic profiles and station-related coefficients. The classification scheme makes use of station coefficients, which were shown to be important indicators of T–S characteristics, and the rotation of characteristic profiles, a procedure closely related to factor rotation in factor analysis.

By applying the ASD technique, we were able to objectively classify the original 95 stations into 4

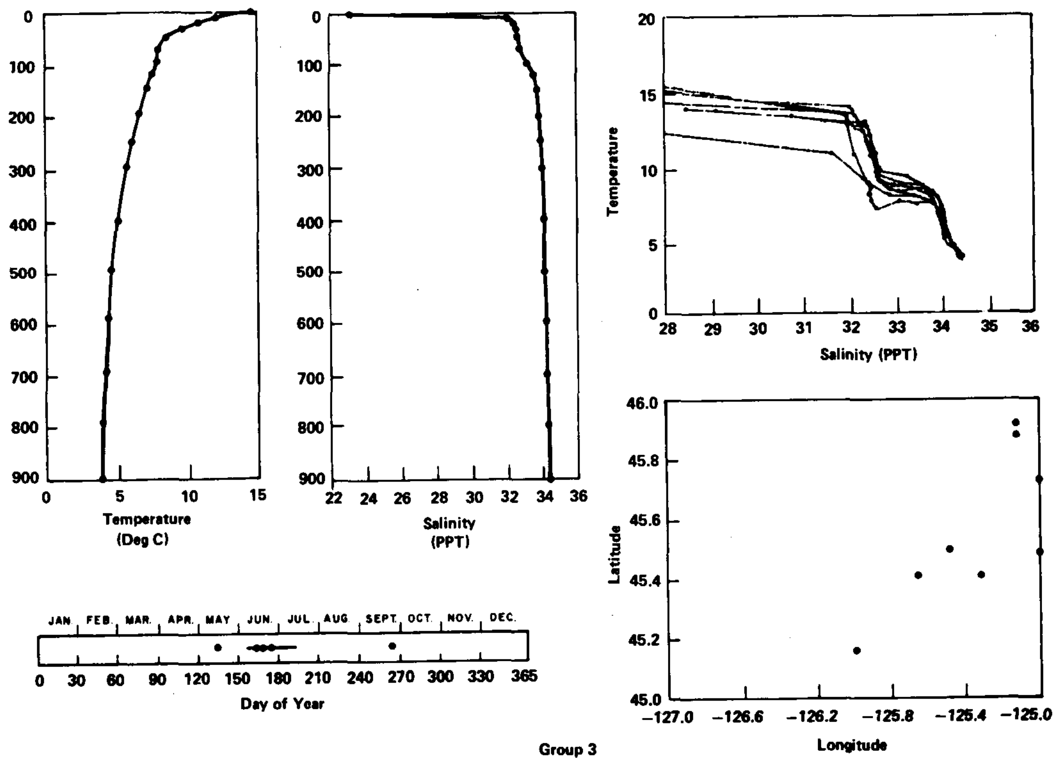


Group 1



Group 2

c



d

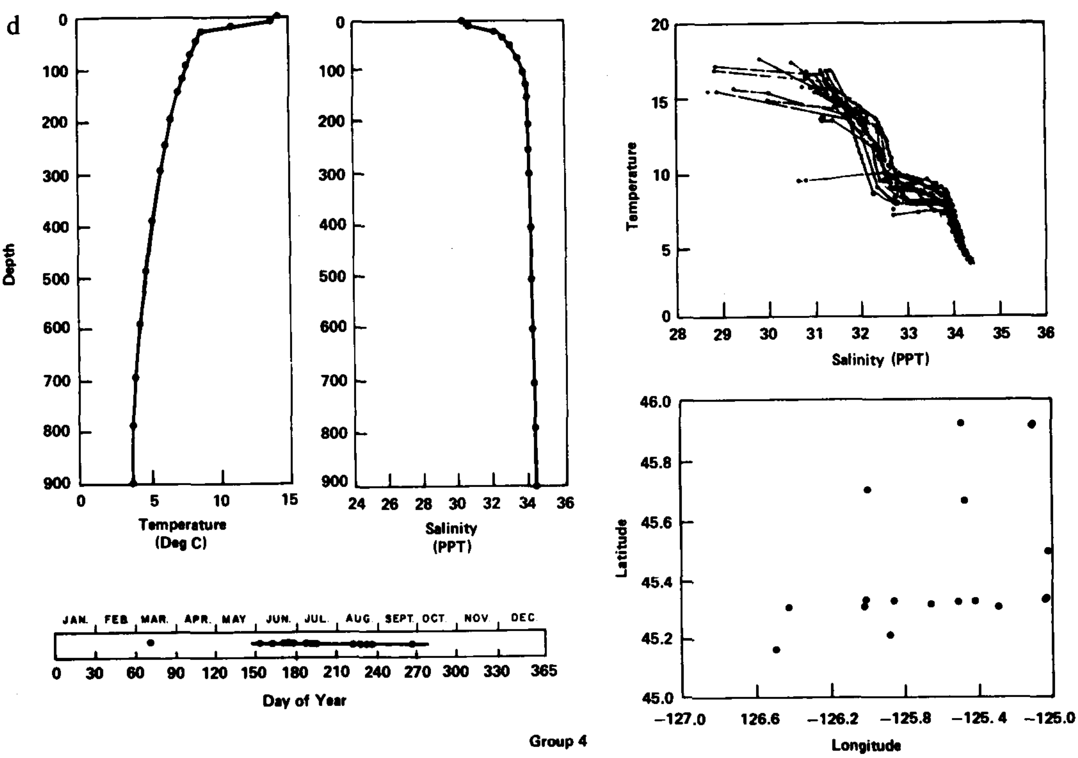


Fig. 10. Results of classification: Group 1 (a), Group 2 (b), Group 3 (c), Group 4 (d).

groups, according to temperature–salinity–depth traits. Furthermore, the groups produced by ASD are supported by environmental factors, such as time, location, and river outflow.

The value of the ASD technique lies in its ability to reduce the original data, consisting of a complex mass of numbers, to a smaller, more important set of numbers. The end result of the analysis, however, depends on the skill with which the products of the technique are assembled in order to uncover significant relationships.

REFERENCES

Barnes, C. A., Duxbury, A. C. and Morse, B. 1972. Circulation and selected properties of the Columbia River effluent at sea. Chapter 3 in *The Columbia River estuary and adjacent ocean waters.* Bioenvironmental Studies (eds. A. T. Pruter and D. L. Alverson). Seattle: University of Washington Press, pp. 41–80.

Dodimead, A. J., Favorite, F. and Hirano, T. 1963. Review of oceanography of the Subarctic Pacific region. *Salmon of the North Pacific Ocean,* Part II. International North Pacific Fisheries Commission, *Bull. No. 13.* Vancouver, Canada.

Good, I. J. 1969. Some applications of the singular decomposition of a matrix. *Technometrics 11* (4), 823–831.

Golub, G. H. and Reinsch, C. 1970. Singular value decomposition and least squares solutions. *Num. Math. 14,* 403–420.

Jalickee, J. B. 1976. Asymptotic singular decomposition of large matrices, in preparation.

Powell, M. J. D. 1964. An efficient method for finding the minimum of a function of several variables without using derivatives. *Comp. J. London* 7 (2), 155–162.

Whittle, P. 1952. On principal components and least square methods of factor analysis. *Skand. Aktuar. 35,* 223–239.

ОБЪЕКТИВНЫЙ АНАЛИЗ И КЛАССИФИКАЦИЯ ОКЕАНОГРАФИЧЕСКИХ ДАННЫХ

Представлен новый подход к анализу и классификации океанографических данных. Методика, основанная на теореме сингулярных разложений, является эмпирической и употребляется для характеристики профилей температуры и солёности с глубиной также, как при анализе водных масс. В дополнение к профилям используются коэффициенты, зависящие от расположения станций и показывающие сходства или различия в соответствии с распределениями станций в пространстве и по времени в наборе данных. Эти коэффициенты служат для классификации отдельных станций по группам со сходными профилями. Представлены результаты применения нового метода к данным океанических станций у побережья штата Орегон.