ABSTRACT

Cluster analysis was applied to XBT profiles in the Indian Ocean in order to describe patterns of ocean variability. Once such patterns were established, maximum likelihood classification was applied to edit the data by assigning the profiles to one of the patterns or rejecting them if they did not meet a certain criterion. The errors of commission and omission were 0.02% and 0.5%, respectively.

METHODS

In order to assess the validity of data going into the model, the data must be separated into subgroups that represent dominant regimes in temperature and salinity. Podeszwa (10) published a report in which he studied sound-speed profiles in the Indian Ocean. The model, which is a four-dimensional steady-state digital model of ocean temperature and salinity (6), requires data which are homogeneous in space and time. The Indian Ocean, being such a highly dynamic system presents problems in meeting these requirements. The driving winds of the southwest (June-September) and northeast (December-February) monsoons are responsible for large fluctuations in the seasonal heat storage across the ocean, particularly in the northwestern Indian Ocean/Arabian Sea areas (4). Vertical and horizontal advection brought on by these winds as well as evaporation, and increased intensity of eddy-field circulation during the monsoons cause large variations in the density structure. The southwest monsoon largely influences the Somali Current which is associated with the cold upwelling off the Somali coast. Heating in the upper 400 m during the northern spring is redistributed later in the summer, with a heat loss in the upper 100 m and heat gain in the 100- to 400-m layer (4).

An example of some of these rapid changes is illustrated by Bruce (2) in figure 1, where XBT profiles were taken approximately at the same location, 12 days apart. Bruce mentions that between 20 August 1979 and 1 September 1979, the 20° isotherm rose 138 m.
Ocean and subdivided the area into 8 profile envelopes or groups. In general, these groups had the same sound-speed profile below 1200 ft. The 8 groups had a varying number of profiles within them. These are shown with letters A through G in figure 2, with the subgroup designated by numbers. These groups or provinces were used as the study units for the deep-ocean GDEM data set (below 400 m). The subgroups were used for the upper model (0-400 m). This grouping satisfies the requirement for space homogeneity because profiles within each group do not differ significantly from each other, whereas profiles between groups do differ (10).

The requirement for homogeneity in time is partially met by subdividing the data into seasons. The seasons were determined as those periods which encompass the extreme effects of the monsoon winds and the periods in between. This determination was made using annual profiles within each province, at 0, 100 and 200 m. Because the model will be a monthly model at the surface, and because below 400 m there is little or no seasonal variability, the data at 100 and 200 m were chosen to determine the seasons. A look at the 100-m data showed extreme variability because of the constantly changing depth of the thermocline. Thus, the data at 200 m were considered the most appropriate for seasonal partitioning. The seasons were chosen as:

- **Spring** - March 1 to mid-May,
- **Summer** - mid-May to mid-July,
- **Fall** - mid-July to mid-October,
- **Winter** - mid-October to March 1.

The deep models will be determined using these seasons and then merged with the top-models at monthly intervals.

The requirement of time stationarity is only partially met by the seasonal partition, therefore, a further partition is required within each season, within each province. Since the data represent a mixture of profiles from different heating, evaporation and circulation patterns, they must be subjected to an analysis which will group similar patterns together. These patterns can then be used as the basis for comparing and editing the individual XBT profiles.

Although some excellent techniques are available for analyzing univariate mixtures of distributions (9), a multivariate technique is desirable because an independent test at each depth is not capable of detecting a profile that oscillates within the envelope.

The multivariate techniques to be used require some prior knowledge about the distribution of the data. Often samples are taken in an environment where the individual distributions are not readily known or available. The sample in figure 1 might be considered a sample from a mixture of 2 p-dimensional distributions. Consider a mixture of c p-dimensional distributions. The probability function for the mixture samples is given as

$$p(x | \Theta) = \sum_{j=1}^{c} p(x | \theta_j, \phi_j) \cdot \phi_j$$

where $x$ is a p-dimensional data vector, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_c\}$, and unknown parameter vectors $\phi_j = \text{class } j, j=1, \ldots, c$.

In order to study each distribution or characteristic profile pattern individually, a method is required to decompose the mixture into its components. Many problems arise in this situation in obtaining a unique solution. (The reader is referred to reference [7] for an in-depth discussion on this topic).

One approach often taken by researchers to describe substructures or components within a data set is cluster analysis. Cluster analysis is not an analytical...
The nearest-neighbor single-linkage technique is an agglomerative clustering method (7). It begins with an n-partition \( C = \{X_1, \ldots, X_n\} \).

Using a similarity criterion, such as distance, the nearest pair of observations is merged. If \( X_i \) and \( X_j \) are merged into \( C_i \), then \( C_j \) is discarded and the mean of \( C_i \) is recomputed. The number of partitions, \( c \), is reduced by 1. The distances are recomputed from the new cluster center to the rest of the clusters and the procedure is repeated until the distance between the nearest pair exceeds a threshold. This threshold may be chosen arbitrarily or based on a realistic requirement.

The distance measure used for this study is a weighted distance given by

\[
d_{ij} = \frac{1}{\sigma_{kk}} \|X_{ik} - X_{jk}\|_2 \quad i = 1, \ldots, n, \quad j = 1, \ldots, i, \quad i \neq j
\]

\( X_{jk} \) the \( k^{th} \) feature of \( j^{th} \) data vector, \( \sigma_{kk} \) the variance of the \( k^{th} \) feature.

When the sample space has been reduced to \( c \)-components, the resulting clusters are used as seed clusters for further clustering.

The technique then employed is a basic minimum squared-error technique where samples are moved from one cluster to another in order to achieve an "optimal" partition based on the minimum squared-error criterion (7). This criterion can be defined as

\[
J_e = \sum_{i=1}^{c} J_i
\]

\[
J_i = \frac{n_i}{n_i + 1} \|X - m_i\|^2
\]

The squared error for the \( j^{th} \) cluster increases to

\[
J_j^* = J_j + \frac{n_j}{n_j + 1} \|X - m_j\|^2
\]

when a new sample \( X \) is added to it (7) and the error for the \( i^{th} \) cluster decreases to

\[
J_i^* = J_i - \frac{n_i}{n_i + 1} \|X - m_i\|^2
\]

when \( X \) is removed from it. If the decrease in \( J_i \) is greater than the increase in \( J_j \), then \( X \) is moved to cluster \( C_j \). This is equivalent to moving \( X \) to cluster \( C_i \) if

\[
\frac{n_j}{n_j + 1} \|X - m_j\|^2 < \frac{n_i}{n_i + 1} \|X - m_i\|^2
\]

The procedure continues until \( J_e \) no longer changes or a certain number of iterations is reached. The problem of estimating the sampling distribution and proving cluster validity statistically is an essentially unsolved problem (7). However, a goodness-of-fit measure is desirable as a test for profiles which fit or do not fit the characteristics of the clustered data. The following approach is used as a test. Once the partitions are defined, a maximum likelihood classifier is employed to assign vectors to one of these partitions using a probabilistic level of confidence. Sample statistics are derived from the \( c \)-components to estimate parameters in \( \theta \). It is assumed, for the purpose of constructing a test, that these components of the mixture are all multivariate normally distributed. The statistics then used are:

\[
\hat{\theta} = \text{p-dimensional sample mean vector}
\]

\[
\hat{\Sigma} = \text{p x p sample covariance matrix}
\]

\( \hat{\theta} \) is always a symmetric, positive semidefinite matrix. Using Bayes rule, a probability statement can be given for each cluster-derived population, \( \omega_j \).

Let \( p(\omega_j|X_i) = p(\omega_j|X_i)p(\omega_j)/p(X_i) \)

\[
p(\omega_j|X_i) = \frac{p(X_i|\omega_j)p(\omega_j)}{\sum_j p(X_i|\omega_j)p(\omega_j)}
\]

be the probability of population \( \omega_j \), given the observation vector \( X_i \); and let \( p(\omega_j) \) be the a priori probability of class \( \omega_j \). Then, if we assume we have normally distributed variables, \( X_i \sim \text{MVN}(\hat{\theta}, \hat{\Sigma}) \), then the probability statement can be computed

\[
p(\omega_j|X_i) = \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2}(X_i - \theta_j)^T \hat{\Sigma}^{-1} (X_i - \theta_j)\right]
\]

\[
p(\omega_j)/p(X_i|\omega_j) p(\omega_j)
\]

The denominator is a constant with respect to \( i \) and need not be computed for assignment purposes. Usually, the log of the numerator is used for computational facility.

The \( \omega_j \) with the highest probability is then chosen as the most likely group for that \( X_i \).
An important factor in the decision-making process is how much error should be tolerated. Because the goal here is to identify data that falls into homogeneous groups and reject data that does not, a criterion is needed. The term \( Y AY \), where \( Y \) is a normally distributed \( p \)-vector, \( A \) is an idempotent, positive definite symmetric matrix of rank \( p \), is distributed as a chi-square, \( \chi^2 \), with \( p \) degrees of freedom \((11)\).

The term, \( (X-\mu)' \begin{bmatrix} I_p & \beta' \end{bmatrix} (X-\mu) \), is therefore distributed as \( \chi^2 \) with \( p \) degrees of freedom, and may be used to test whether an observation is an acceptable candidate for membership in one of the distributions. The choice of the acceptable error in rejecting a good observation is arbitrary. Usually, this error, called a type-I error (or omission error) is preferable to a type-II error (or commission error), the acceptance of a bad observation. Therefore, \( \alpha \), the probability of a type-I error, is chosen to minimize the error which is most costly.

**RESULTS**

The data used are from the Navy Master Oceanographic Data Set (MOODS), produced by Compass Systems, Inc. for the Fleet Numerical Oceanography Center. Data were interpolated to 0, 100, 200, 400, 1000, 2000, 3000, and 4000 m for this particular study. The analysis for the top model was based on data at 0, 100, 200, and 400 m, \( p=4 \), and for the bottom, all data below 400 m, \( p=5 \). These depths are used only for this analysis and not for the actual model building.

Figure 4 shows an edited spring data set for province B3, which is also in the monsoon-affected area. The figure shows the union of all the data that were accepted into one of the clusters. Figure 5 shows the set of rejected clusters. Figure 6 shows a set of data classified into 1 cluster.

An example of the cluster-analysis output is shown in figure 3. Each one of these profiles represents an average of all profiles in that cluster. The dashed line represents the mean of all the data for that cluster analysis. This figure shows an area, province B82, off the Somali coast, where upwelling is very prominent. This particular set of clusters comes from the month of August when the southwest monsoon is generally at its peak.

Although the complete set of results from this study have not been compiled, a preliminary set shows the rejection rate to be between 3 and 11% of the data, the average being about 6%. This rate is somewhat high in some cases due to the fact that there is
often no data at the top or bottom of the profile. Since clusters are derived only from profiles for which all the data is present, incomplete profiles must be classified using a partitioned mean and covariance matrix. Profiles with missing data at the surface are not used. Many profiles which are sampled to only 100 or 200 m do not classify into the established clusters, and these cause the rejection rate to be as high as 11% in some cases, although they only account for a fraction of that 11%.

The rejection of "good" profiles, or omission error, is about 0.5%. This figure is somewhat subjectively determined, as it is computed on what one thinks is a good profile. The commission error, or acceptance of bad data, is determined on the basis of those profiles which were thought to be bad. This number was determined to be 0.02%.

SUMMARY

The feasibility of using cluster analysis and maximum likelihood classification for data editing in the Indian Ocean has been studied and appears to give reliable results. A few problems exist with the cluster analysis as far as the stability of the cluster means, variance and number of clusters, given changes in the merging distance. When the distance is increased, variances increase, therefore increasing the noisiness of the final output data. These changes did not seem to affect the final outcome so much when there was an adequately large sample. With smaller samples, it is often necessary to merge data in adjoining provinces to be able to cluster samples without having to greatly increase the merging distance. This merging usually had to be done at depths below 1000 m because of inadequate sampling. Because of the stability of temperatures at these depths throughout the year, and because only one season was used below 400 m, this merging was only needed a few times, and seemed to have no effect on the outcome.

The method which has been advocated depends on some assumptions about the normality of the cluster samples. The low error rates indicate that the technique is not very sensitive to deviations from normality that might be present in the samples. Most of the problems which would result from such deviations would appear in borderline cases because of tight constraints on the merging distance. Since cluster centers are close to the mode, severe deviations from normality would cause a rejection of values in the long tails of a skewed distribution. If the tails are long, the data would cluster if there were a large proportion of data there, otherwise they would be rejected. By keeping the merging distance small, problems with accepting data in nonexistent tails is avoided. Commission error rates indicate that the technique is not overly sensitive to these problems. Therefore, it is recommended as a data-editing tool for highly variable data sets such as the Indian Ocean.

REFERENCES


