J. Paleont., 67(3), 1993, pp. 475–486

0022-3360/93/0067-0475$03.00

ERROR-WEIGHTED MAXIMUM LIKELIHOOD (EWML): A NEW STATISTICALLY BASED METHOD TO CLUSTER QUANTITATIVE MICROPALEONTOLOGICAL DATA

EVAN FISHBEIN AND R. TIMOTHY PATTERSON

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, 911091 and
Ottawa-Carleton Geoscience Center and Department of Earth Sciences, Carleton University,
Ottawa, Ontario, K1S 5B6, Canada

ABSTRACT—The advent of readily available computer-based clustering packages has created some controversy in the micropaleontological community concerning the use and interpretation of computer-based biofacies discrimination. This is because dramatically different results can be obtained depending on methodology. The analysis of various clustering techniques reveals that, in most instances, no statistical hypothesis is contained in the clustering model and no basis exists for accepting one biofacies partitioning over another. Furthermore, most techniques do not consider standard error in species abundances and generate results that are not statistically relevant. When many rare species are present, statistically insignificant differences in rare species can accumulate and overshadow the significant differences in the major species, leading to biofacies containing members having little in common.

A statistically based "error-weighted maximum likelihood" (EWML) clustering method is described that determines biofacies by assuming that samples from a common biofacies are normally distributed. Species variability is weighted to be inversely proportional to measurement uncertainty. The method has been applied to samples collected from the Fraser River Delta marsh and shows that five distinct biofacies can be resolved in the data. Similar results were obtained from readily available packages when the data set was preprocessed to reduce the number of degrees of freedom. Based on the sample results from the new algorithm, and on tests using a representative micropaleontological data set, a more conventional iterative processing method is recommended. This method, although not statistical in nature, produces similar results to EWML (not commercially available yet) with readily available analysis packages. Finally, some of the more common clustering techniques are discussed and strategies for their proper utilization are recommended.

INTRODUCTION

MICROPALEONTOLOGISTS OFTEN use quantitative analysis of fossil faunas to determine the number and characteristics of different environments (biofacies) represented by the samples under consideration. When the number of samples is small, it is possible to intuitively define the biofacies boundaries and subdivide samples. However, if the number of samples is large, or if the differences between biofacies are subtle, more rigorous analytical methods are required to distinguish discrete environments. In these cases, computer-based multivariate analysis tools must be used to determine the relationships among samples.

Until recently, this sort of analysis was impossible for the scientist unversed in computer programming and statistical methods (Hooper, 1969a, 1969b; Yzerdraat et al., 1969; Buzas, 1970, 1979). A survey of paleontological studies from the 1950's, 1960's, and even the early 1970's shows that very few researchers utilized computers or multivariate analysis to classify their data. However, with the recent proliferation of microcomputers and the accompanying rapid development of off-the-shelf statistical programs, most micropaleontologists now utilize some form of computer-based multivariate analysis, ranging from cluster analysis to principal component analysis, in their research. Unfortunately, many paleontologists have only a limited background in statistics and often use techniques that are inappropriate to the underlying statistical hypotheses. This problem arises because many popular software packages have not been specifically designed to perform the biofacies analysis required by the paleontologist and many of the subroutines contained in these programs utilize algorithms with little significance to paleontological applications.

The purpose of this paper is to present a new statistically significant "error-weighted maximum likelihood" (EWML) method of clustering. Since the EWML clustering method is not presently commercially available, a procedure for obtaining similar results with off-the-shelf software is presented. Presently available methods of cluster analysis are evaluated in the Appendix. For those not familiar with this form of multivariate analysis, the underlying concepts are explained and defined. This information will facilitate the choice of clustering software and development of analysis strategies most likely to produce reliable results.

A STATISTICALLY DERIVED CLUSTERING METHOD

The basic objective of a cluster analysis is to determine the number of clusters represented by a data set and to determine the probable affinity of the various component samples. Clustering packages contain many clustering strategy options (similarity, linkage, and clustering algorithm). Dramatically different results can be obtained from the same data set depending on the clustering strategy chosen (see Appendix). In most cases, deciding on the best clustering strategy requires some statistical hypothesis to define a biofacies. Unfortunately, the commonly available strategies lack any statistical basis and all results are therefore equally valid or invalid (Buzas, 1979).

The difficulty in formulating a clustering algorithm based on statistical hypotheses arises from the absence of a mathematical definition of a biofacies. Intuitively we know that samples from a biofacies have similar characteristics, although some variability is expected. This similarity is mathematically described by the average fractional abundance or centroid. The centroid is "on average" the expected abundances in the samples (see Appendix). Variability is best explained by example. Consider a hypothetical biofacies containing just two species (species 1, species 2). One hundred samples are collected and their fractional abundances are plotted in Figures 1 and 2. Assuming normal and uniform distribution in Figure 1 the sum of the two abundances adds to 1, so the samples lie along a line. The centroid, plotted as a +, lies towards the center of the line.

The standard deviation σ, shown by the box, characterizes variability around the centroid and marks the biofacies boundary. However, σ does not uniquely characterize variability. For example, Figure 2 is derived from a second biofacies having the

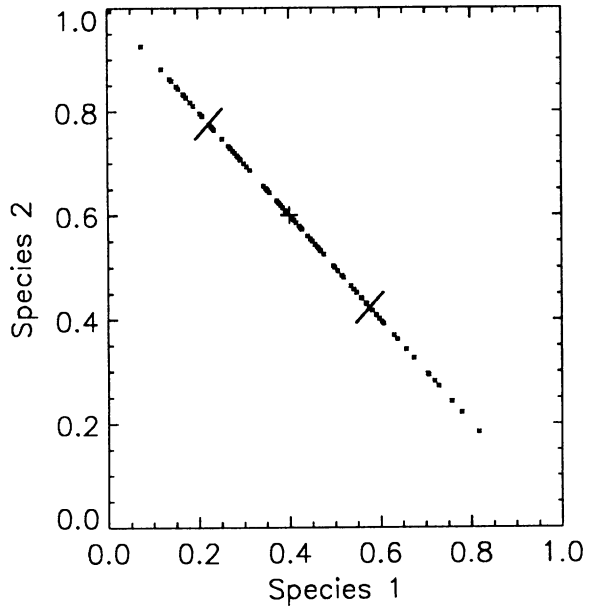

FIGURE 1—Sample scatter diagram of hypothetical species 1 versus species 2 plane. Points are the positions of 100 samples. The plot is a straight line because each sample contains only these two species (summing to one). The centroid (+) is plotted toward the center of the line. The data points enclosed by the two lines perpendicular to the axis of plotted points represent the biofacies boundary, which characterizes the variability (standard deviation) around the centroid.

same centroid and standard deviation, but having samples concentrated near the edges of the boundary of the box. Additional parameters are required to mathematically distinguish the two biofacies and quantify variability.

Biofacies containing more than two species require multidimensional statistical methods, but the concepts illustrated by the previous example are the same. For example, consider a biofacies having a third species collected in two hundred samples (Figure 3). The centroid still characterizes the average sample, but the variability and biofacies boundaries are indicated by an ellipse. The variability is now calculated from the covariance matrix. This has important consequences when ascertaining if a sample belongs to a particular biofacies. Variability is most pronounced in one direction, as shown by the elongation of the ellipsoid. For example, the average species distribution within the ellipse is 30 percent species 1, 50 percent species 2, and 20 percent species 3. Sample 1, with 48 percent, 39 percent, 13 percent of species 1, 2, and 3, respectively (Figure 3), is inside the biofacies boundary. The Euclidean distance (see Appendix for definition) from the centroid is 0.22. Meanwhile sample 2, with abundances 30 percent, 50 percent, 20 percent, and separated from the centroid by a distance of only 0.09, is outside the boundary and statistically less likely to be from this biofacies, even though it is geometrically closer to the average. Most currently available clustering methods would measure the Euclidean distance from the centroid and would consider sample 2 to be more representative of the biofacies than sample 1 in the biofacies. This is despite the graphical evidence that shows that the first sample differs from the centroid in a fashion more typical of the biofacies. For many distributions, specifying the centroid, standard deviation, and form of the covariance matrix completely describes the biofacies statistically.

The "error weighted maximum likelihood" (EWML) procedure contains the statistical hypothesis that samples from similar environments (biofacies) are distributed normally around a centroid, and that a covariance matrix describes the amount of variability of samples within the biofacies.

For any sample, the fractional abundance of the ith species has an uncertainty (Dx_i) because it is estimated from a sample containing a finite number of specimens (Patterson and Fishbein, 1989). Consequently, although a sample was introduced as a point in abundance space, it is actually a volume (ΔV) where $\Delta V = \Delta x_1, \Delta x_2 \ldots \Delta x_K$ surrounding the abundance (x_i). Samples with abundances that are contained within the uncertainty volume of other samples are statistically indistinguishable. A successful clustering algorithm should not distinguish between samples having no statistical differences.

A statistical model containing these criteria is expressed in equation 1. For a sample s(m) derived from a biofacies which varies normally around average values $\vec{\mu}$, the density f(m) of samples in the volume ΔV centered at x(m) is the likelihood of finding a sample in this region. When the density is one, any possible sample one might collect would have abundances somewhere in the volume. Likewise when the density is near zero, so is the fraction of samples contained in ΔV.

$$f(x(m)) = \frac{1}{(2\Pi)^{k/2}|S|^{1/2}} \int_{x_1(m)-\Delta x_1(m)}^{x_1(m)+\Delta x_1(m)} \cdots \cdot \int_{x_K(m)-\Delta x_K(m)}^{x_K(m)+\Delta x_K(m)} \exp\left\{-\frac{1}{2}\sum_{r=1}^{K}\sum_{s=1}^{K} S^{-1}{}_{rs}(X'_r - \mu_r)(X'_s - \mu_s)\right\} \cdot dx'_1 \ldots dx'_K. \quad (1)$$

It is therefore most reasonable to conclude that a sample having these abundances is from another biofacies.

The right side of equation 1 is a statement of the statistical hypothesis that samples are normally distributed around the centroid and that the covariance matrix S or its inverse characterizes the width of the variability. The argument of the exponential function

$$\sum_{r=1}^{K}\sum_{s=1}^{K} S^{-1}{}_{rs}(x'_r - \mu_r)(x'_s - \mu_s) \quad (2)$$

is the squared Mahalanobis distance (defined in the Appendix) between the centroid and any arbitrary point x′ with coordinates x'_r or x'_s. Species "t" which vary greatly have small coefficients $S^{-1}{}_{tt}$. The term in the sum where r, s = t, $S^{-1}{}_{tt}(x'_t - \mu_t)^2$ has a relatively small contribution to the distance unless $x'_t - \mu_t$ is large. Likewise species "v" which have small variability will have a large $S^{-1}{}_{vv}$ and the term $S^{-1}{}_{vv}(x'_v - \mu_v)^2$ will have a large contribution to the sum, unless x'_v is almost equal to μ_v. Similarly, if two species "t, v" have correlated variabilities, the coefficient $S^{-1}{}_{vt} = S^{-1}{}_{tv}$ will be large or small depending on whether the correlation is small or large.

The factor $-\frac{1}{2}$ multiplying the squared Mahalanobis distance gives the distribution its maximum at the centroid while normalizing the width of the distribution to the covariance matrix described below. The exponential function is one when its argument is zero and decreases rapidly to zero when its argument is greater than 1 (Mahalanobis distance greater than 2). The biofacies boundary occurs at a Mahalanobis distance of 1.

The density function is averaged over the uncertainty of the measurements produced by statistical counting errors (Patterson and Fishbein, 1989). The integrals over each of the coordinates centered on the sample abundances accomplish this averaging. Lastly, the factor $1/(2\Pi)^{k/2}|S|^{1/2}$, where $|S|$ is the determinant of S, normalizes the exponential so that the total density over

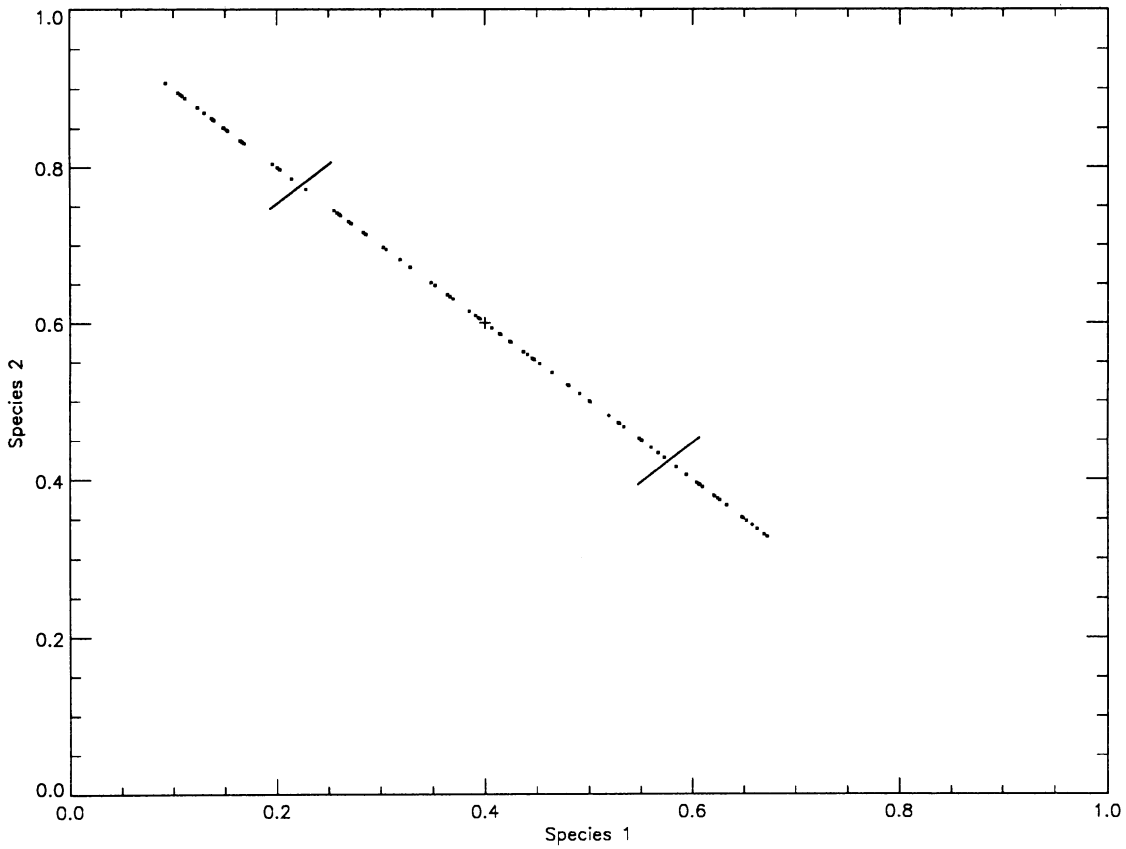

FIGURE 2—Sample scatter diagram of hypothetical species 1 versus species 2 plane. Points are the positions of 100 samples. The plot is a straight line because each sample contains only these two species (summing to one). The centroid (+) is plotted toward the center of the line. The data points enclosed by the two lines perpendicular to the axis of plotted points represent the biofacies boundary, which characterizes the variability (standard deviation) around the centroid. The centroid and standard deviation illustrated by this biofacies is the same as shown in Figure 1 but the samples are concentrated near the edges of the biofacies boundary.

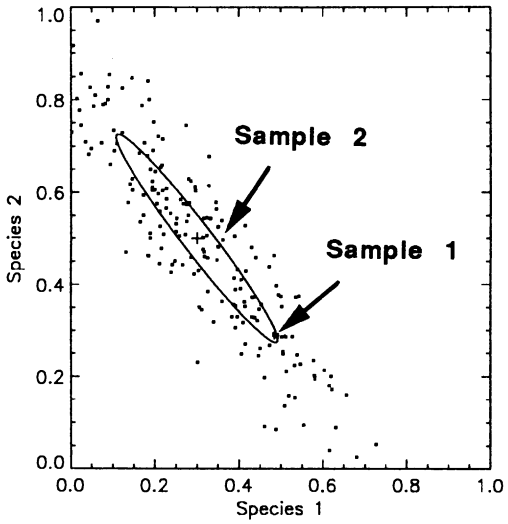

FIGURE 3—Sample scatter diagram of hypothetical species 1 versus species 2 plane. Points are the positions of 200 samples. The centroid (+) is plotted toward the center of the ellipse. The ellipse surrounding many of the data points represents the biofacies boundary and characterizes the variability (standard deviation) around the centroid. The projections of the standard ellipsoids for the biofacies characterized by these species is shown from upper left to lower right. Variability is primarily in one direction, as shown by the elongation of the ellipsoid. Sample 1 is inside the biofacies boundary (Euclidean distance

all possible species abundances is 1. This normalization is valid only when the variables (abundances) can be arbitrarily negative or positive. An assumption is made that at the extremes of valid abundances (0, +1) the Mahalanobis distance is much larger than one.

The covariance matrix and its inverse are symmetrical matrices that stretch and rotate lines radiating from the centroid. A set of "K" directions exists (see Cushing, 1975) that the covariance matrix only stretches. The directions are always mutually perpendicular. The K directions are given the symbols $\vec{e}$ . . . $\vec{e}_k$ and the amount of stretching is given symbols (σ_1 . . . σ_k). The original lines $\vec{e}$ . . . $\vec{e}_k$ radiating from the centroid define a sphere. Transformation of those directions by the covariance matrix produces lines $\sigma_1\vec{e}_1$ . . . $\sigma_k\vec{e}_k$ that define an ellipsoid.

← = 0.22) but sample 2, although closer to the centroid (Euclidean distance = 0.09), is statistically less likely to be from this biofacies. Most available clustering methods using Euclidean distance from the centroid as the primary determinant of similarity would place the second sample in the biofacies before the first. However, the graphical evidence shows that the first sample differs from the centroid in a fashion more typical of the biofacies. For example, the average species distribution is 30 percent species 1, 50 percent species 2, and 20 percent species 3. For many distributions, specifying the centroid, standard deviation, and form of the distribution function completely describes the biofacies statistically.

The product of the axes lengths (s_1, s_2 . . . s_k), is proportional to the volume of the standard ellipsoid and is also the square root of the determinant (S). Equation 1 is therefore the ratio of the volume of the uncertainty divided by the volume of the standard ellipsoid, times a weighting function that depends only on the Mahalanobis Euclidean distance to the centroid.

As stated above, the principal ellipsoid is the surface where the Mahalanobis distance is equal to 1. Solving for this surface is an essential part of the EWML clustering method. The covariance matrix, set of directions $\vec{e}$ . . . $\vec{e}_k$ and stretching (σ_1 . . . σ_k) all have K(k + 1)/2 independent coefficients (symmetry reduces the number from K^2) because they are orthogonal. The covariance matrix can be written in terms of σ_i and $\vec{e}_i$.

$$S_{rs} = \sum_{k=1}^{K} \sigma_k{}^2 e_{rk} e_{sk}, \qquad \text{where r, s} = 1 \ldots K. \tag{3}$$

$$S^{-1}{}_{rs} = \sum_{k-1}^{K} \sigma_k{}^{-2} e_{rk} e_{sk}, \qquad \text{where r, s} = \ldots K. \tag{4}$$

Equation 3 is a set of K(k + 1)/2 independent linear equations for finding the directions $\vec{e}_i$ and the amount of stretching. The probability is constant whenever the argument of the exponential is constant.

Substituting equation 3 into equation 2, the Mahalanobis distance is 1 and probability is constant when the Euclidian distance σ_n from the centroid along $\vec{n}$ is

$$\sigma_n = \left\{ \sum_{r=1}^{K} \frac{1}{\sigma_r{}^2} \left(\sum_{L=1}^{K} n_i e_{ri} \right)^2 \right\}^{\frac{-1}{2}} \tag{5}$$

where σ_n is the radius from the centroid to the surface. This was the basic equation used to graph the principal ellipsoid in any arbitrary plane and was used for Figures 1 and 2. The innermost sum in equation 5 is the dot product or direction cosines between $\vec{n}$ and $\vec{e}_r$.

Any biofacies determination should include examination of the correlation matrix's standard ellipsoid. The standard ellipsoid, which is a surface of constant probability, provides a picture of the distribution of samples in the biofacies. In directions where the ellipsoid is thin, the biofacies have little variability. Similarly, in directions in which the ellipsoid is elongated, fractional abundances have great variability. Often the axes are rotated between species abundance directions. This means that the abundance of one species varies in proportion to variations in the abundances of other species. This variability could be characteristic of a species' sensitivity to environmental variability. For example, within a salt marsh, abundances of halophilic and halophobic foraminifera would be inversely correlated, owing to salinity variability. This would be represented by some of the axes having components simultaneously along halophilic and halophobic species directions. However, some correlation arises because the sum of fractional abundances is equal to 1. For example, the biofacies from a salt marsh containing only two species, both halophilic, will show an inverse correlation and have the standard ellipse inclined 45° from the species directions. Examining the shape and orientation of the principal ellipsoid assists the paleontologist in determining species relationships, in assessing when biofacies can be associated with a particular environment, and if biofacies variability is normally distributed (described by equation 1).

Solving equation 2 is often referred to as the "principal component" or "eigenvalue problem." The vectors ($\vec{e}_k$), referred to as the principal components, represent a linear combination of the observed variables, provide maximal discrimination of the samples under study, and do not correlate with any other principal components. The choice of coordinate system is arbitrary. The most obvious set is species direction. For statistical analysis, it is convenient to define the "principal coordinate system" (PCS), which has its origin at the center of the distribution. Coordinates are measured along the axes of the principal ellipse. The principal coordinate system is specific to a biofacies. The introduction of PCS allows considerable simplification of the basic equation (equation 1). The transformation from species axes to PCS axes involves a translation ($x - \mu$) and a rotation $\vec{e}_{ri}$ between old and new coordinate directions. In the principal coordinate system, each sample has coordinates

$$y_i(m) = \sum_{r=1}^{K} (e_{ri} x_r - \mu_r), \qquad i = 1 \ldots K \tag{6}$$

and uncertainties

$$\Delta y_i{}^2(m) = \sum_{r=1}^{K} S_{ri}{}^2 x_r{}^2, \qquad i = 1 \ldots K. \tag{7}$$

In the principal coordinate system, the probability function (equation 1) reduces to the particularly simple form

$$f(m) = \prod_{i=1}^{K} \left\{ \frac{1}{(2\pi)^{1/2}\sigma_i} \int_{y_i(m)-\Delta y_i(m)/2}^{y_i(m)+\Delta y_i(m)/2} \exp \frac{-y_i{}^2}{2\sigma_i{}^2} \, dy'_i \right\} \tag{8}$$

in which the volume integral ln(1) converts into a product of independent one-dimensional integrals (the symbol X means to multiply all terms with i [i = 1, 2, . . . K]).

The uncertainties of a sample's fractional abundances are often greater than the Euclidean distance from the centroid. This is usually true if some of the species are rare. In such cases, some of a sample's principal coordinates will be smaller than the principal coordinate uncertainties. For those principal coordinates, the one-dimensional probability integral in equation 8 (the quantity between brackets) is approximately equal to 1. Axes of the principal ellipse for which this is true are called "unresolved axes." The value of these principal coordinates can have any magnitude smaller than the principal uncertainty without changing the probability.

Determining the probability that a sample is derived from a particular environment requires estimates of both the centroid and covariance matrix. An unbiased estimate of the centroid based on (M) samples is given by

$$\mu_i = \frac{1}{M} \sum_{m=1}^{M} x_i(m). \tag{9}$$

The centroid given by equation (9) is not equivalent to the mean used by Buzas (1990), in which specimens of all samples within a biofacies are accumulated into a single "super sample." This is equivalent to an error-weighted average. Intrinsic to this formulation is the notion that variation within a biofacies is smaller than counting errors. Using his suggestion when the biofacies variability is large will distort the estimate of the centroid towards samples with better statistics (larger specimen counts). For example, with reference to Figure 2, if the samples in the upper left corner contain more specimens than samples in the lower right corner, application of an error-weighted average will push the estimate toward the upper left corner. Clearly the differences are artifacts of the counting method. Buzas's mean will only be equivalent to the centroid if an equal number of specimens is counted in each sample, a generally unpractical requirement (Figure 2).

The probability that a sample is derived from a biofacies exists, even when its determinant is zero and the inverse of the covariance matrix does not exist. Wilks (1962) provided an unbiased estimate of the determinant of the covariance matrix (equation 9),

$$|S| = \frac{(m - K - 1)!}{(M - 1)!}\,|u|, \tag{10}$$

where u is the scatter matrix defined in equation 20. An estimate of the covariance matrix consistent with equation 10 is given in equation 11.

$$S_{ij} = \left(\frac{(M - K - 1)!}{(M - 1)!}\right)^{1/K} \sum_{m=1}^{M} (x_i(m) - \mu_i)(x_j(m) - \mu_j). \tag{11}$$

Equations 10 and 11 show that when the number of samples from the biofacies is less than the number of species (K > M), the determinant of the estimated covariance matrix is zero. Since the covariance matrix has an inverse if and only if its determinant is nonzero, it is not obvious that equations 1 or 8 can be evaluated. Analysis of the principal component problem of the scatter matrix (u_{ij}) shows that some of the axes (K + 1 − M axes) have a zero length. These zero length axes, called "undetermined axes," arise because the (M) data points are described by M − 1 perpendicular vectors. Even though some of the principal coordinates are undetermined, the expansion of the sample's fractional abundances as presented in principal components is not arbitrary because these principal components are always zero. Examining equation 8 in the limit where a principal component has a zero length shows that the term in brackets is equal to one.

The process of gathering more information about the biofacies reduces the probability when there are undetermined axes. The K + 1 − M undetermined axes exist because insufficient information is provided to estimate them. The zero length is an assumed default length in the absence of more data. It is also assumed that any length, including zero, has significant implications for paleontological analysis. A length equal to zero implies that variability along that direction is much smaller than the uncertainty of the fractional abundances. In this case probability is improved by uncertainty and clustering favors clusters having fewer samples than species. The optimum clustering consists of clusters containing a single sample.

There are also large uncertainties in estimates of the probability when some of the principal ellipsoid axes are undetermined. If undetermined axes are assumed to be large, then fractional abundances are evenly distributed. In this case it can be shown that for a large number of samples the centroid is located at a value of 1/2, and the axes have a length of $12^{-1/2}$ (29 percent). Assuming the value of undetermined axes to be 29 percent, the bracketed term in equation 6 has a smaller value equal to $\Delta y/(0.29\sqrt{2\pi})$.

For a hypothetical cluster, a within-cluster similarity measure can be developed that maximizes the probability that a collection of samples is derived from a single biofacies. By substituting equations 9, 10, and 11 into equations 6 and 8, the within-cluster similarity (for a single cluster) becomes

$$\ln(f^{-1}) = \frac{K_a}{2}\left(\frac{(M - 1)!}{(M - K_a - 1)!} + \frac{M}{2}\ln(2\pi)\right) + \frac{M}{2}\ln(|S|) - \sum_{m=1}^{M}\sum_{i=1}^{K_a} \ln(\Delta y_i(m)). \tag{12}$$

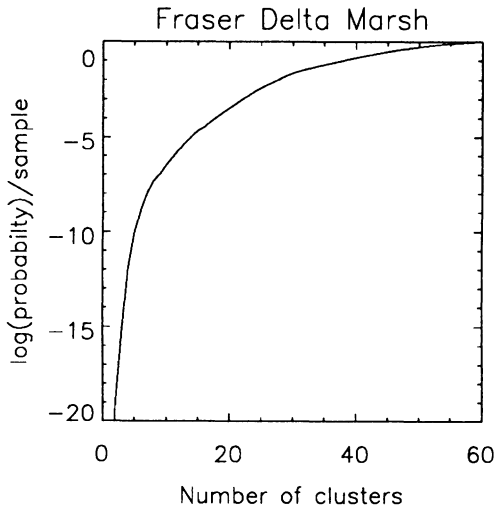

FIGURE 4—Averaged natural logarithm of sample probability for the most probable n-cluster clustering versus the number of clusters.

The determinant of the covariance matrix in equation 12 is obtained by assuming values for the undetermined axes and by removing (integrating over y_i) those axes that are unresolved. A principal axis smaller than the mean principal uncertainty (Δy_i) for all samples is treated as if it were unresolved. Samples are averaged over the resolved axes and indexed from 1 to K_a, where K_a is the number of resolved axes. The integrals in equation 8 were evaluated by assuming that the projected abundance uncertainties were either much smaller or much larger than the length of the axes. For biofacies determination, the similarity measure described by equation 12 has four desirable features: 1) biofacies similarities, smaller than abundance uncertainties, are removed from the similarity measure; 2) samples are assumed to have ellipsoidal symmetry, which allows for biofacies containing correlated variability of species abundances; 3) biofacies centroids and covariance matrices are calculated, providing a more complete characterization of the biofacies than sample membership alone; and 4) the probability of assembling a cluster is calculated, providing a mechanism for comparing clusterings not having the same number of biofacies.

BIOFACIES DETERMINATION OF A PALEONTOLOGICAL DATA SET FROM THE FRASER DELTA, BRITISH COLUMBIA

Clustering using the EWML method provides information not available from clustering programs in off-the-shelf statistical packages and eliminates spurious correlations. To demonstrate the procedure, biofacies were determined for foraminifera-bearing samples collected from modern marshes fringing the Fraser River Delta. The data set consists of 60 samples, in which a total of 17 species of foraminifera were found (see Patterson, 1990, for detailed tabulation of the species present). As demonstrated elsewhere (Scott and Medioli, 1980; Scott, 1976; Goldstein and Frey, 1986), the foraminifera found in marshes can be subdivided into distinct biofacies, correlating with such parameters as differences in elevation, salinity, and organic content of the surficial sediments. Because of the small number of taxa generally encountered and the distinctiveness of the various subenvironments of the marsh, marsh biofacies can often be recognized without recourse to computers. For these reasons, samples from the Fraser River Delta marsh provide an ideal data set to demonstrate the utility of this new clustering technique.

The Fraser Delta samples were clustered using a hierarchical

TABLE 1—Characteristics of the five biofacies represented in Fraser Delta Marsh samples. Included are: number of samples; location and standard error of the centroid; directions cosines; and length and uncertainty of length of the standard ellipsoid's axes.

		Species									
		A. beccarii	*C. gunteri*	*J. macrescens*	*M. fusca*	*T. inflata*	*T. pacifica*	*A. exiguus*	*H. advenum*	*A. salsum*	Axis length
Biofacies A	mean	2.7 ± 2%	0.0%	0.1 ± 0.1%	77.0 ± 0.4%	0.0%	0.0%	17.9 ± 0.4%	0.0%	2.2 ± 0.2%	
18 samples	axis	0.012	0.000	−0.001	−0.741	0.000	0.000	0.669	0.000	0.060	25.7 ± 1.6%
	direc-	0.795	0.005	0.002	−0.396	0.000	0.000	−0.457	0.000	0.048	4.3 ± 1.2%
	tion	−0.340	0.004	0.013	−0.218	0.000	0.000	−0.312	0.000	0.860	2.2 ± 0.9%
Biofacies B	mean	0.1 ± 0.2%	0.0%	71.9 ± 0.7%	4.1 ± 0.4%	22.7 ± 0.7%	0.0%	0.0%	1.2 ± 0.3%	0.0%	
14 samples	axis	0.000	0.000	0.730	−0.047	−0.682	0.000	0.000	−0.002	0.000	28.2 ± 2.6%
	direc-	−0.021	0.000	0.454	−0.631	0.531	0.000	0.000	−0.338	0.000	5.0 ± 2.1%
	tion	−0.074	0.000	0.136	0.604	0.106	0.000	0.000	−0.775	0.000	2.0 ± 1.3%
Biofacies C	mean	69.6 ± 1.0%	0.0%	1.7 ± 0.3%	25.8 ± 1.0%	0.1 ± 0.3%	0.0%	2.5 ± 0.4%	0.2 ± 0.3%	0.1 ± 0.3%	
12 samples	axis	−0.721	0.000	−0.037	0.687	0.000	0.000	0.080	−0.007	0.001	31.7 ± 3.5%
	direc-	0.423	0.000	−0.810	0.403	−0.015	0.000	−0.020	−0.019	0.015	3.7 ± 2.3%
	tion	0.192	0.000	0.278	0.318	0.062	0.000	−0.880	−0.008	0.037	1.0 ± 1.9%
		0.233	0.000	0.248	0.227	−0.029	0.000	−0.845	0.199	−0.241	0.5 ± 1.5%
Biofacies D	mean	0.0%	0.0%	0.0%	2.8 ± 0.8%	0.0%	95.9 ± 0.9%	0.6 ± 0.7%	0.6 ± 0.6%	0.2 ± 0.6%	
4 samples	axis	0.000	0.000	0.000	−0.790	0.000	0.587	−0.001	0.172	0.032	5.1 ± 1.7%
	direc-	0.000	0.000	0.000	−0.239	0.000	−0.213	0.875	−0.357	−0.066	1.7 ± 1.4%
	tion	0.000	0.000	0.000	−0.149	0.000	−0.412	0.154	0.796	−0.389	1.2 ± 1.4%
Biofacies E	mean	17.8 ± 1.3%	11.1 ± 0.6%	0.1 ± 0.3%	63.3 ± 0.4%	0.7 ± 0.5%	0.0%	4.7 ± 0.5%	0.0%	2.1 ± 0.4%	
4 samples	axis	0.182	−0.687	−0.007	0.680	0.068	0.000	−0.111	0.000	−0.127	20.4 ± 2.1%
	direc-	0.478	−0.417	−0.007	−0.444	−0.177	0.000	0.604	0.000	−0.062	6.1 ± 1.9%
	tion	0.497	0.354	0.071	0.052	0.015	0.000	−0.185	0.000	−0.763	1.3 ± 1.5%

algorithm and a linkage based on the within-cluster similarity linkage described by equation 12 (EWML method). In this configuration the unresolved standard ellipsoid axes are assumed equal to zero. The average natural logarithm of the probability per sample for the most probable hierarchical n-cluster clustering is graphically presented (Figure 4). The monotonic decrease in probability results from the assumption that undetermined axes of the standard ellipsoid are equal to zero. The logarithm of the probability decreases at a uniform rate between 60 and eight clusters (potential biofacies), decreases at a slightly greater rate between eight and four clusters, and plummets when the number of clusters is fewer than four. The rapid decrease for fewer than four clusters indicates that within-cluster variability is decreasing at a rate too large to be attributed simply to unresolved axes. Therefore, using the result presented in this graph, it can be concluded that between four and eight biofacies are represented by these samples.

Table 1 summarizes the results of a cluster analysis that produces seven biofacies. The centroids indicate that only four or five of the 17 species (dimensions) present in abundance space are significant in each biofacies and only nine species (accounting for greater than 98.8 percent of the specimens) warranted inclusion in the table. This is not surprising since every species is absent in at least one sample, and most species are absent in most samples. Three of the clusters contain 73 percent of the samples while the remaining four clusters each contain four samples.

Of the four clusters having four samples, two have axes lengths greater than 65 percent abundance. The orientation of these two axes indicated that these clusters were distinct from the remaining clusters. However, the length of the axes was much larger than the width of a uniformly distributed variable (29 percent), indicating that these two clusters are not statistically significant. They therefore have not been included in Table 1 or in the discussion of the analysis.

Three of the clusters (A, B, C) contained at least 12 samples each. In the absence of species abundance correlation, this is enough samples to determine 11 axes of the standard ellipsoid. However, two clusters are three dimensional and the other is four dimensional. The dimensionality of the ellipsoid is always smaller than the number of species present in the biofacies. The uncertainty associated with the locations of the centroids indicates that Clusters A and B contain four primary species while Cluster C has five. Therefore, Clusters A and B should have ellipsoids that have three axes easily resolved and Cluster C should have four. The low dimensionality indicates complete partitioning of some species to some biofacies and poorly resolved fractional abundances for most species.

Correlation between species is represented by large eccentricity, when the axes do not lie along species abundance directions. For example, in Cluster C (Biofacies C) the orientation of the largest axis indicates that most of the variability occurs as an exchange between *Ammonia beccarii* (Linné, 1758) and *Miliammina fusca* Brady (*in* Brady and Robertson, 1870). This correlation is most easily seen in a scatter plot (Figure 5) on the *Ammonia beccarii* versus *Miliammina fusca* plane. The ellipse obtained by cutting a slice through the standard ellipsoid on the *Ammonia beccarii* versus *Miliammina fusca* plane is obtained with equation 6. The eccentricity of the ellipse is 0.999 and the major axis is inclined −45° from the vertical (negative correlation). The eccentricity is close to one because 95.4 percent of the specimens are *Ammonia beccarii* plus *Miliammina fusca* (the direction of the longest axis) and variability of this line is only 4.6 percent (1.00–95.4 percent). Because Biofacies C is composed primarily of two species, the variability within the biofacies contains induced correlation produced by the nor-

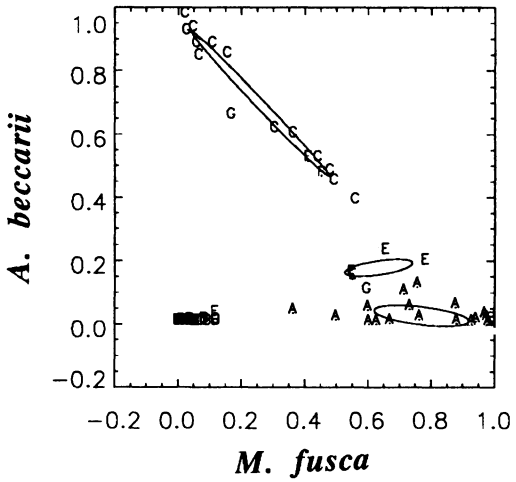

FIGURE 5—Sample scatter diagram on the *Ammonia beccarii* versus *Miliammina fusca* plane. Letter indices are the positions of associated biofacies. The projections of the standard ellipsoids for Biofacies C, E, and A are shown from upper left to lower right. Biofacies B is not represented in here because there are almost no *Ammonia beccarii* or *Miliammina fusca*. The axes of the standard ellipsoid are reported as direction cosines (from the species directions) and the axis length as percentage abundance. For all clusters, the eccentricity is greater than 0.94. However, since the eccentricity of a sphere and straight line are 0.0 and 1.0, respectively, an eccentricity close to 1.0 suggests that any clustering algorithm assuming spherical clusters (e.g., based on an unnormalized Euclidean distance) would give spurious results.

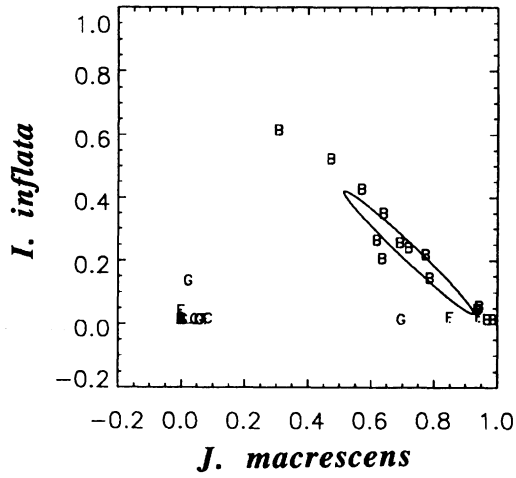

FIGURE 6—Sample scatter diagram on the *Trochammina inflata* versus *Jadammina macrescens* plane. Letter indices are the positions of associated biofacies. The projection of Biofacies B appears at the left of the figure. The standard ellipsoids of Biofacies A, C, D, and E project as horizontal lines or points at the origin. The axes of the standard ellipsoid are reported as direction cosines (from the species directions) and the axis length as percentage abundance. For all clusters, the eccentricity is greater than 0.94. However, since the eccentricity of a sphere and straight line are 0.0 and 1.0, respectively, an eccentricity close to 1.0 suggests that any clustering algorithm assuming spherical clusters (e.g., based on an unnormalized Euclidean distance) would give spurious results.

malization of fractional abundances and cannot be biologically interpreted.

Biofacies A is also composed primarily of two species, *Miliammina fusca* and *Ammobaculites exiguus* Cushman and Brönniman, 1948. The largest axis is inclined 45° from the abundance directions, reflecting the fact that the sum of the abundances of *Miliammina fusca* and *Ammobaculites exiguus* equals 1. Figure 5 shows that Biofacies A has a weak negative correlation between a rare species (*Ammonia beccarii*) and a common one (*Miliammina fusca*). The variation in *Ammonia beccarii* is indicated by the second largest axis of the standard ellipsoid. While this axis is much smaller than the first, its error is small compared to its length (Table 1), indicating that biological significance can be attached to the negative correlation between *Ammonia beccarii* and *Miliammina fusca*. *Ammonia beccarii* is characteristic of the higher low marsh (Biofacies C), while *Miliammina fusca* predominates in Biofacies A in the lower low marsh. Moving from the higher low marsh to the lower low marsh, the proportions of these dominant species covary with an increasing proportion of *Miliammina fusca* at lower elevations.

Biofacies E contains four samples, has a considerable amount of scatter, and would not normally be considered adequately sampled. However, all of the samples containing *Cribroelphidium gunteri* (Cole, 1931), an abundant species, cluster in this biofacies. In addition, all of these samples are contained within the higher low marsh zone, indicating some elevational control for this biofacies. The primary variability within this environment occurs primarily in the *Cribroelphidium gunteri* and *Miliammina fusca* direction, with significant variability in the *Ammonia beccarii, Ammobaculites exiguus,* and *Ammotium salsum* (Cushman and Brönniman, 1948) directions. While most of the sample populations do not consist entirely of two species, variability occurs predominantly as an exchange between *Cribroelphidium gunteri* and *Miliammina fusca*. Biofacies E is unusual; although *Ammonia beccarii* and *Miliammina fusca* are both common, their correlation is weakly positive, inclined at +80° from the vertical. If more samples were available, this correlation could prove biologically significant.

Biofacies B contains 14 samples that cluster along a line inclined 44° from the horizontal in Figure 6, indicating composition primarily of two species, *Trochammina inflata* (Montagu, 1808) and *Jadammina macrescens* (Brady *in* Brady and Robertson, 1870). It is notable that these two species are common only in Biofacies B. Both of these species are characteristic of high marsh environments. Biofacies B, of this study, contains the *Jadammina macrescens–Trochammina inflata* Biofacies as recognized by Patterson (1990). Patterson's biofacies were distinguished based on nonstatistical criteria established by Scott (1976). The present study suggests that this distinction is not statistically valid.

Biofacies D, containing four samples, is composed primarily of *Trochammina pacifica* (Cushman, 1925), and has only 5.1 percent species variability. The standard ellipsoid appears in Figure 5 as a short horizontal line near the origin. The variability is along a direction approximately in the *Miliammina fusca* versus *Trochammina pacifica* plane, but has a significant component in the *Haplophragmoides manilaensis* (Andersen, 1953) direction. The samples bearing *Trochammina pacifica* were all from a sheltered area between two causeways. The restriction of this species to this area is probably related to the high organic content of the substrate (Patterson, 1990).

Obtaining similar results using off-the-shelf software.—As the EWML method of clustering described above is not presently available commercially, the previous discussion will be of limited utility to micropaleontologists who are neither mathematically inclined nor skilled in programming. It seemed useful to determine a combination of more common methodologies that would best emulate the statistically valid results of EWML. Several commercially available procedures were applied to the same data set in an attempt to produce comparable results.

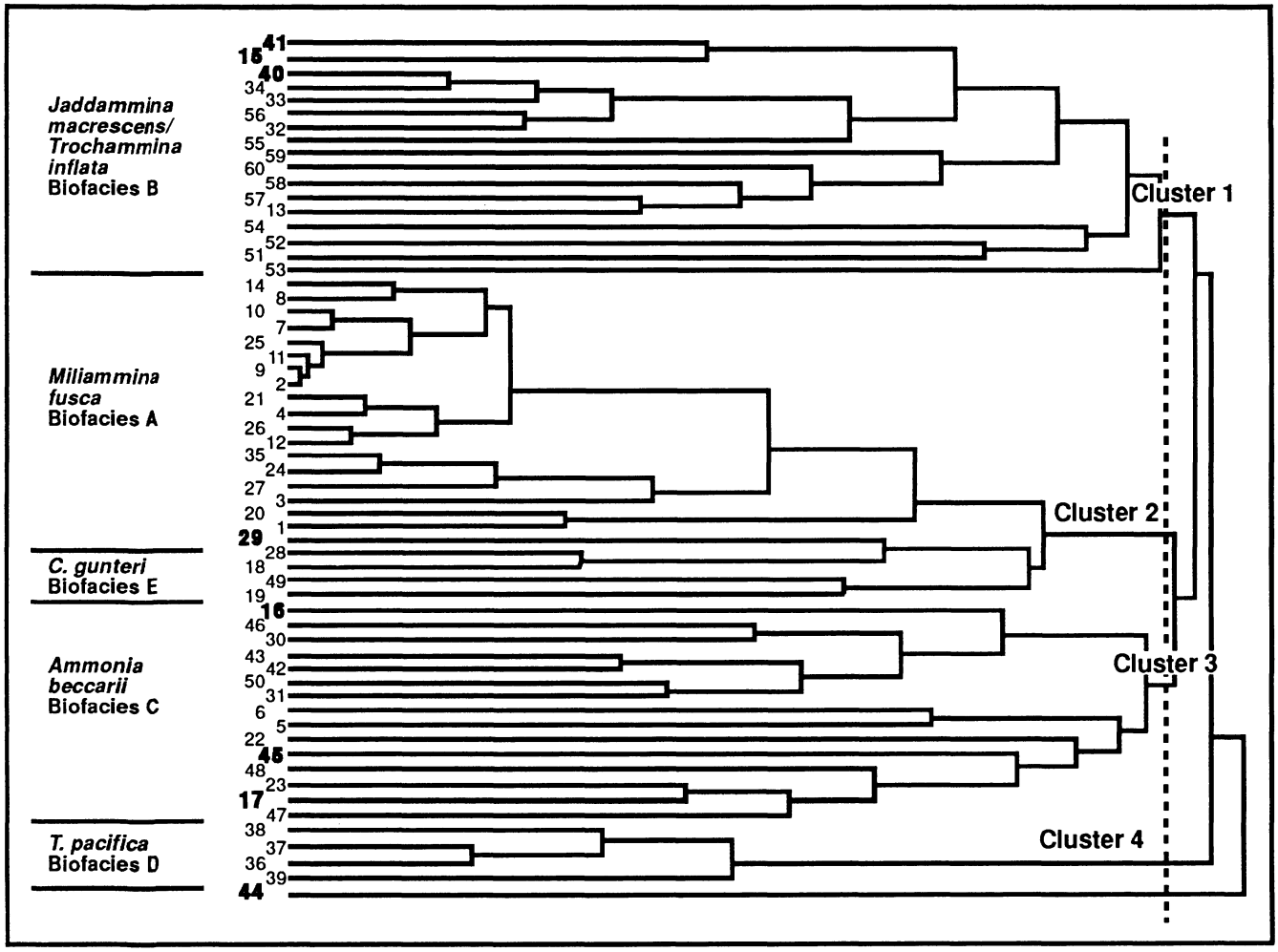

FIGURE 7—Q-Mode dendrogram of Fraser Delta data set generated by clustering first four principal components using average linkage. Four major clusters (biofacies) are recognized with this method, indicated by the dotted line, as opposed to the five biofacies discerned using EWML (listed at the left). Samples listed in bold type could not be resolved by EWML. Sample 44 could not be resolved adequately by either EWML or the SYSTAT statistical package.

All of the biofacies, as defined by EWML, have very elongated and standard ellipsoids. A consideration of the discussion presented earlier in the paper makes it clear that clustering algorithms based on Mahalanobis Euclidean distance, using a pooled-within covariance matrix or the covariance matrix for the entire data set, will not produce paleontologically reasonable clusters.

The best result obtained by any procedure first required considerable condensation of the data to remove unnecessary "noise." Instead of the 17 species analyzed using EWML, a reduced data set containing most of the information in the original data with a fewer number of degrees of freedom (numbers of species) was derived. Furthermore, species were eliminated that strongly covaried with other species. For example, the relative frequencies of *Ammobaculites exiguus, Ammotium salsum,* and *Miliammina fusca* covaried significantly in most samples. Thus, the statistics associated with *Ammobaculites exiguus, Ammotium salsum,* and *Miliammina fusca* were merged into a single variable that contained this relationship. The data set was condensed by performing a principal component analysis on the raw data and rotating the raw data onto the principal components. Principal component analysis provides a powerful tool for analyzing the differences between biofacies. This procedure is equivalent to treating the data set as a single biofacies, calculating the principal ellipsoid for the entire data set, and finding each sample's principal coordinates (solving equations 5 and 6). The rotated data set was condensed by eliminating those principal coordinates having relatively small lengths. These first four principal components explained 97.77 percent of the data. Following this examination, a Q-Mode hierarchical cluster analysis was carried out using the first four principal components, a Euclidean similarity measure, and average linkage (Figure 7). It is of interest that for this particular data set, almost identical results were obtained using a centroid linkage. Distinct clusters of samples with average similarities greater than an arbitrarily selected level were considered biofacies. Scatter plots of the entire set of samples should be made on the planes spanned by the two largest axes of the standard ellipsoid of each cluster. The scatter plots should be examined to see if the distribution of points belonging to the cluster associated with the plane of the scatter plot are distributed around a centroid, and if points from other clusters are separated from the cluster associated with the plane of the scatter plot. Finally, the scatter plots produce a visual tool for determining if the number of clusters should be changed.

Clusters 1, 3, and 4, as generated by the statistics package, were almost identical to Biofacies B, C, and D produced by

EWML. The commercially available statistics package could not resolve the difference between Biofacies A and E, lumping these groups in Cluster 2. It has already been pointed out that this grouping, characterized by the presence of *Cribroelphidium gunteri,* would not normally be considered to have been sampled adequately. Biofacies A and E both have high proportions of *Miliammina fusca,* which explains why the statistics package clustered them together.

Many biofacies, such as those defined by benthic foraminifera from sets of samples derived from deep-sea-drilling cores, contain hundreds of species. Typically all but a few species are rare (Boltovskoy, 1978). Based on this example, hundreds of samples and tens of thousands of counts per sample would be required to resolve a large fraction of the standard ellipsoid's axes for all species in such samples. Therefore, considerable forethought as to which species are essential to define a biofacies and the number of counts necessary (Patterson and Fishbein, 1989) is required before carrying out a cluster analysis using the recommended off-the-shelf methodology described above.

SUMMARY OF RECOMMENDED CLUSTERING PROCEDURES

A new statistically valid "error-weighted maximum likelihood" (EWML) method of clustering paleontological data sets has been presented. However, as this method is not yet commercially available, a procedure for obtaining similar, though not statistically significant, results using available statistical packages has been devised.

The recommended strategy for data analysis is: to cluster between cases (samples), not variables (species); to use species fractional abundances in the data array; and to reduce the dimensionality and eliminate similarity in rare species with unresolved fractional abundances by 1) eliminating statistically insignificant species from the matrix; 2) applying a standard principal component analysis to the scatter matrix; 3) calculating the mean fractional abundance uncertainty,

$$\frac{1}{M}\left\{\sum_{m=1}^{M}\sum_{i=1}^{K} C_i^2(m)\right\}^{1/2};$$

and 4) keeping principal coordinate directions having lengths greater than the fractional abundance uncertainty. In addition, data should be clustered with a hierarchical algorithm, 1) using an unnormalized Euclidean or squared Euclidean distance with either a complete, average linkage or Ward's linkage (see Appendix); and 2) determining the cluster boundaries by subjectively cutting the graphically displayed branches in hierarchical dendrograms at an intuitive level of similarity (see Figure 7).

Ways to enhance the reliability of the clustering process include: 1) reducing the number of clusters by combining clusters when the centroid of one lies within the standard ellipsoid of another; 2) performing K-means clustering to refine hierarchical clustering using the previous results to specify the number of clusters; and 3) completing analysis by calculating centroids and performing principal component analyses on each cluster.

ACKNOWLEDGMENTS

Acknowledgment is made to the National Research Council of the National Academy of Sciences for support of this research to E.F. as a Resident Research Associate at the Jet Propulsion Laboratory. This research was also partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Operating Grant OGPOO41665 to R.T.P. The research described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We thank H. E. Anderson, M. A. Buzas, and C. T. Schafer for critically reviewing the manuscript.

REFERENCES

ABRAMOWITZ, M. A., AND I. A. STEGUN. 1972. Handbook of Mathematical Functions with Formulae, Graphs and Mathematical Tables. U.S. Government Printing Office, Washington, D.C., 1,046 p.

ANDERBERG, M. R. 1973. Cluster Analysis for Applications. Academic Press, New York, 359 p.

ANDERSEN, H. V. 1953. Two new species of *Haplophragmoides* from the Louisiana Coast. Contributions from the Cushman Foundation for Foraminiferal Research, 4:20–22.

BOLTOVSKOY, E. 1978. Late Cenozoic benthonic foraminifera of the Ninetyeast Ridge (Indian Ocean). Marine Geology, 26:139–175.

BRADY, G. S., AND D. ROBERTSON. 1870. The Ostracoda and Foraminifera of tidal rivers with an analysis and description of the Foraminifera. Annual Magazine of Natural History, 6:273–309.

BUZAS, M. A. 1970. On the quantification of biofacies. Proceedings of the North American Paleontological Convention, Part B:101–116.

——. 1979. Quantitative biofacies analysis. Foraminiferal Ecology and Paleoecology. SEPM Short Course No. 6:11–20.

——. 1990. Another look at confidence limits for species proportions. Journal of Paleontology, 64:842–843.

COLE, W. S. 1931. The Pliocene and Pleistocene foraminifera of Florida. Bulletin of the Florida State Geological Survey, 6:7–79.

CRONBACH, L. J., AND G. C. GLESER. 1953. Assessing the similarity between profiles. Psychological Bulletin, 50:456–473.

CUSHING, J. T. 1975. Applied Analytical Mathematics for Physical Scientists. John Wiley & Sons, New York, 651 p.

CUSHMAN, J. A. 1925. Recent foraminifera from British Columbia. Contributions from the Cushman Laboratory for Foraminiferal Research, 1:38–47.

——, AND P. BRÖNNIMAN. 1948. Additional new species of arenaceous foraminifera from the shallow waters of Trinidad. Contributions from the Cushman Laboratory for Foraminiferal Research, 24:37–43.

FRIEDMAN, H. P., AND J. RUBIN. 1967. On some invariant criteria for grouping data. Journal of the American Statistical Association, 62: 1159–1178.

GOLDSTEIN, S. T., AND R. W. FREY. 1986. Salt marsh foraminifera, Sapelo Island, Georgia. Senckenbergiana Maritima, 18:97–121.

HARTIGAN, J. A. 1975. Clustering Algorithms. John Wiley & Sons, New York, 351 p.

——, AND M. A. WONG. 1979. A K-means clustering algorithm: algorithm AS 136. Applied Statistics, 28:456–473.

HOOPER, K. 1969a. Processing of foraminiferal data: a computer program, p. 291–306. *In* P. Brönniman and H. H. Renz (eds.), Proceedings of the First International Conference on Planktonic Microfossils, Vol. II, Geneva, 1967. E. J. Brill, Leiden.

——. 1969b. A re-evaluation of eastern Mediterranean foraminifera using factor-vector analysis. Contributions from the Cushman Foundation for Foraminiferal Research, 20:147–151.

JARDINE, C. J., N. JARDINE, AND C. SIBSON. 1967. The structure and constitution of taxonomic hierarchies. Mathematical Biosciences, 1:173–179.

JOHNSON, S. C. 1967. Hierarchical clustering schemes. Psychometrika, 32:241–254.

LANCE, G. N., AND W. T. WILLIAMS. 1967. A general theory of classificatory sorting strategies. 1. Hierarchical systems. Computer Journal, 9:373–380.

LINNÉ, C. 1758. Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cim characteribus, differentiis, synonymis, locis. G. Engelmann, Lipsiae, ed., 1:1–824.

MONTAGU, G. 1808. Testacea Britannica, supplement. Exeter, England. Printed by S. Woolmer, 183 p.

PATTERSON, R. T. 1990. Intertidal benthic foraminiferal biofacies on the Fraser River Delta, British Columbia: modern distribution and paleoecological importance. Micropaleontology, 36:229–244.

——, AND E. FISHBEIN. 1989. Re-examination of the statistical methods used to determine the number of point counts needed for micropaleontological quantitative research. Journal of Paleontology, 63: 245–248.

SCOTT, D. B. 1976. Quantitative studies of marsh foraminiferal pat-

terns in southern California and their application to Holocene stratigraphic problems, p. 153–170. *In* C. T. Schafer and B. R. Pelletier (eds.), First International Symposium on Benthonic Foraminifera on Continental Margins, Part A, Ecology and Biology. Maritime Sediments, Special Publication 1.

——, AND F. S. MEDIOLI. 1980. Quantitative studies of marsh foraminiferal distributions in Nova Scotia. Implications for sea level studies. Cushman Foundation for Foraminiferal Research Special Publication No. 17, 58 p.

SNEATH, P. H. A., AND R. R. SOKAL. 1973. Principles of Numerical Taxonomy. W. H. Freeman, New York, 574 p.

WARD, J. H., JR. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58:236–244.

——, AND M. E. HOOK. 1963. Application of an hierarchical grouping procedure to a problem of grouping profiles. Educational and Psychological Measurement, 23:69–82.

WILKS, S. S. 1960. Multidimensional statistical scatter, p. 486–503. *In* I. Olkin (ed.), Contributions to Probability and Statistics. Stanford University Press.

——. 1962. Mathematical Statistics. John Wiley & Sons, New York, 644 p.

YZERDRAAT, W., K. HOOPER, AND B.-D. ERDTMANN. 1969. Fortran programs for faunal analysis. Carleton University, Department of Geology Geological Paper 69-3, Ottawa, Canada, 106 p.

ZAHN, C. T. 1971. Graph-theoretical methods for detecting and describing Gestalt clusters. IEEE Transactions on Computers, C-20:68–86.

ACCEPTED 5 MAY 1992

APPENDIX

COMMON CLUSTERING TECHNIQUES

Clustering by sample, also known as Q-Mode Analysis, refers to the division of samples into disjoint sets. Each cluster represents samples from a single environment. Several basic components, including the similarity measure, the type of linkage used, and the partitioning algorithm, are needed to define the clustering scheme. However, depending on how these components are implemented, radically different clusters may result. The following section provides definitions of the various components of a cluster analysis and outlines some of the more common methodologies, with particular emphasis on those applicable to paleontological biofacies discrimination.

Fractional abundance.—Each sample is characterized by the various fractional abundances of the component species. The fractional abundance is the number of specimens of each species divided by the total number of specimens. The fractional abundance [x_i (m)] of species (i) in sample (m) is

$$x_i = \frac{C_i(m)}{N(m)} \tag{13}$$

where [C_i (m)] is the number of counts of species (i) in sample (m) and [N(m)] is the total number of specimens of all species (K species are present). The fractional abundance of all species in a sample is therefore

$$N(m) = \sum_{i=1}^{K} C_i(m). \tag{14}$$

A sample's fractional abundances are not independent since they always sum to one. Samples are considered to be similar if the differences in their fractional abundances are small. The set of a sample's fractional abundances can be visualized as the coordinates of a point in a K-dimensional "abundance space" where each axis measures the fractional abundance of a single species (i). The term cluster analysis arises because samples from a similar environment tend to plot close together, forming a clump or cluster of points.

Similarity.—The arrangement of samples into biofacies requires a measure to determine whether samples cluster together or not. This process requires introducing a quantity called the "similarity" or "association," which measures the differences between samples (Anderberg, 1973). In cluster analysis, similarity is usually defined in terms of a distance between points; e.g., samples with similar fractional abundances have small distances (high degree of similarity) between them. There are several ways of calculating distance between samples, not all of which are appropriate for paleontological applications.

Many statistical packages define similarity in terms of angles between coordinates of the sample and some other fixed point, [$x_i(0)$], usually the origin [$x_i(0) = 0$], or ratios of coordinates. Examples of these approaches (Cronbach and Gleser, 1953; Hartigan, 1975) are the correlation measure [R(x(n), x(m))] between points x(n) and x(m)

$$R[x(n), x(m)] = \sum_{i=1}^{K} |x_i(n) - x_i(0)|\,|x_i(m) - x_i(0)| \tag{15}$$

and the covariance measure C[x(n), x(m)]

$$C[x(n), x(m)] = \frac{R[x(n), x(m)]}{d_E[x(n), x(0)]\,d_E[x(m), x(0)]} \tag{16}$$

where d_E is the Euclidean distance defined in equation (17). These methods of measuring similarity (distance) are useful in determining characteristics of taxa in numerical taxonomy (Sneath and Sokal, 1973). However, they do not really take into account the similar fractional abundances of samples from a common environment and therefore are not particularly useful for biofacies discrimination.

Categorical variables (chi square, phi square, Goodman-Crustal index; see Anderberg, 1973, for definitions) are another class of distance measure. These measures determine the joint occurrences of other variables and are therefore formulated in terms of counts. Unfortunately, these counts are not related to species abundances but to the number of times a hypothesis is satisfied. These methodologies should be avoided unless the statistical hypothesis or environmental affinity are known.

The most commonly used measure is Euclidean distance

$$d_E[x(n), x(m)] = \left\{\sum_{i=1}^{K} (x_i(n) - x_i(m))^2\right\}^{1/2}. \tag{17}$$

The Euclidean distance between similar samples is always small and positive. Furthermore, Euclidean distance is the best way to differentiate biofacies because it is related to the normal distribution, which can describe how samples are distributed within a biofacies. However, problems arise using this method because all species abundances contribute equally to the sample discrimination. This is because small variations of rare species abundances often characterize biofacies affinity better than do larger differences in more common species; the contribution of environmentally important rare species may become lost in the noise of more common taxa. Several variations on Euclidean distance measure have been developed that address this problem, although not all are suited to the problem at hand.

One variation based on Euclidean distance is known as Minkowski distance

$$d_M[x(n), x(m)] = \left\{\sum_{i=1}^{K} (x_i(n) - x_i(m))^p\right\}^{1/p}. \tag{18}$$

Minkowski distance is a generalization of the Euclidean distance containing a free parameter (p) that exaggerates similarity when p is large, but becomes reduced to Euclidean distance when p is equal to 2. The Minkowski distance can improve clustering results when samples from an environment are preferentially distributed near or away from the environmental mean in a nonlinear relation. However, no method is available for estimating the exponent p. Therefore the Minkowski distance should be used with caution.

Another variation on Euclidean distance, the Mahalanobis Euclidean distance (MED),

$$d'_E[x(n), x(m)] = \left\{\sum_{i=1}^{K}\sum_{j=1}^{K} S^{-1}{}_{ij}(x_i(n) - x_i(m))(x_j(n) - x_j(m))\right\}^{1/2}, \tag{19}$$

allows all species to contribute to assessment of biofacies affinity because

distances are weighted proportionately to the expected variability, the covariance matrix S. The weighting appears through the inverse of the covariance matrix S^{-1}.

Often a data set does not allow estimation of the covariance. In such cases the matrix can be completed only if assumptions relating coefficients are introduced. For example, if it is assumed that all the species are unrelated then the off-diagonal coefficients are zero. The number of coefficients is much smaller and more readily estimated from the data. In this case, the covariance matrix becomes the variance matrix. Because species fractional abundances are always dependent within a sample and are often correlated within an environment, use of a MED measure in conjunction with a variance matrix is not recommended.

Mahalanobis Euclidean distance can also be expressed in terms of a covariance matrix. A covariance matrix can be estimated from the entire data set (many biofacies), individual biofacies, or from some average of many biofacies. If a covariance matrix is estimated from the entire data set, then the separation between clusters is simultaneously reduced, along with the normalization of rare and common species abundances. Furthermore, this method assumes that species are simultaneously rare or abundant in all environments. For these reasons a MED measure based on a covariance matrix for the entire data set prevents a clustering algorithm from distinguishing dissimilar samples (Hartigan, 1975).

Normalization can also be based on the covariance matrix for each biofacies (within-cluster covariance). This method is problematic because the composition of the clusters must be known prior to calculating distances. Since the partitioning is determined from the distances, the problem becomes cyclical and requires repeated iterations for solution. When an environment is represented by fewer than K (the number of species) samples, a serious problem arises because the estimate of covariance matrix has no inverse. Distances can be calculated, but only if enough assumptions are introduced to make the covariance matrix nonsingular. It is usually difficult to use a within-cluster covariance matrix in paleontological studies because the numbers of species and samples are not usually comparable. However, because unrelated environments tend to have a markedly different variability, clustering using a Mahalanobis distance containing a within-cluster covariance matrix has the best chance of discriminating biofacies.

The singularity contained in the inverse of the within-cluster covariance can be eliminated by averaging all of the within-cluster covariance matrices to produce a "pooled-within" covariance matrix (Friedman and Rubin, 1967). This eliminates multiple distances between samples and tends to keep clusters separated. Unfortunately, when clusters have similar covariance matrices, the average is not particularly effective at normalizing any single cluster. This situation commonly arises in paleontological studies where a species may be rare in one biofacies but common in another. Weighted distances based on pooled-within covariance matrices therefore only should be used when biofacies are expected to have a similar variability.

Linkage.—"Linkage" or "Clustering Criterion" is a quantity that determines whether a hypothetical partitioning of samples is reasonable (Anderberg, 1973). Linkage can be classified further as "between-cluster linkages" (quantifies comparisons between clusters) or "within-cluster linkages" (quantifies all samples contained within a single cluster).

Extremely fast and computer efficient algorithms, most of which are not particularly useful to micropaleontologists, have been developed using between-cluster linkages. The very popular "single linkage" (Zahn, 1971; Johnson, 1967) is the similarity measure between the two closest points not in the same cluster. Because the structure of the linkage does not guarantee that all members of a cluster are close (Jardine et al., 1967), single linkage often produces long clusters in which non-nearest neighbors are far apart. Therefore, single linkage is not recommended for use in biofacies discrimination. In contrast, "complete linkage," another very common linkage, defines the linkage between clusters as the similarity between the two furthermost points. Theorems from graph theory (Johnson, 1967) have shown that members of clusters formed using complete linkages are maximally connected.

Many between-cluster linkages compare all pairs of points in two clusters. One example is "average linkage," in which the linkage is the mean of all dissimilarity measures between pairs of points lying in different clusters. Centroid linkage (Lance and Williams, 1967) measures the similarity between the centroids of clusters. These methods produce similar results because both disfavor joining clusters with centroids that are far removed from one another. Centroid linkage sometimes produces disjoint clusters (Anderberg, 1973) and is therefore less preferable than average linkage measure.

Within-cluster linkages mimic the biologically reasonable hypothesis that samples from common environments are similar. Ward (1963) and Ward and Hook (1963) introduced a linkage (now known as Ward's linkage) equal to the mean similarity between all samples in a cluster and the cluster's centroid. Ward's linkage is characterized by the property that when an environment has very little variability its linkage is short. When used in conjunction with a Euclidean distance, Ward's linkage is the mean radius of the cluster. However, Ward's linkage should never be used in conjunction with a within-cluster Mahalanobis Euclidean distance based on a within-cluster covariance matrix because the normalized mean radius is always one. Within-cluster linkages can also be formulated without a distance measure. Wilks (1960) and Friedman and Rubin (1967) developed clustering techniques using the magnitude of the determinant of the scatter matrix

$$u_{ij} = \sum_{m=i}^{M} (x_i(m) - \mu_i)(x_j(m) - \mu_j). \quad (20)$$

This linkage is related to the volume of the cluster, and can be short even when distances between members of the cluster are large, provided that some of the variability is correlated. However, this type of similarity is difficult to implement in biofacies discrimination because of two sources of artificial correlation: one induced by normalization to fractional abundances, and a second more important source arising when the scatter matrix is estimated by too few samples. The sources of this artificial correlation is fractional abundance, which always contains some correlation because of its normalization. However, a more important source of artificial correlation arises when the scatter matrix is estimated by too few samples. In such cases, not all of the components of the scatter matrix can be resolved. Because this method of similarity measure is numerically cumbersome, difficult to implement, and often singular, it has been used infrequently. However, this linkage has one major advantage over the previously described linkages, in that it can be derived from a statistical hypothesis. Clustering methods based on the scatter matrix will be explored more fully later.

Clustering.—The methodology or algorithm that uses a linkage to divide a set of samples into subgroups or "clusters" is called clustering. Clustering is accomplished either by minimizing the within-cluster linkage or by maximizing the between-cluster linkage. The simplest optimizing algorithms compare all possible arrangements to find the true optimum. Unfortunately, the number of comparisons is too large (Abramowitz and Stegun, 1972) to be handled, even by the largest supercomputer, for assemblages of more than 15 samples. More pragmatic, computationally possible clustering algorithms are only capable of approximating an optimal solution within a limited set of possibilities.

The most common class of clustering algorithms is referred to as "joining" or "hierarchical" clustering. These terms are derived from the (hierarchical) process (Johnson, 1967) by which clusters are built by merging (joining) pairs of clusters from previous iterations. These methods are characterized by tree diagrams relating pairs of clusters to a parent cluster.

If the variability within the biofacies is much smaller than the differences between biofacies, an optimal number of clusters can be estimated from the relative change in similarity along a path from trunk to tip. When this condition is satisfied, any branch joining clusters from different biofacies will have a discontinuously larger length than any of its branches. This is a useful methodology for determining the boundary between biofacies.

A second class of clustering algorithms, termed "exchange" or "sorting" methods, is useful for improving existing cluster divisions. This methodology begins with a preset number of clusters and proceeds to move individual samples between clusters, preserving the number of clusters while reducing the within-clusters linkage. Unlike hierarchical methods, solutions very close to the true optimum may be found with enough iterations. However, this methodology is limited because given a poor initial guess (the number of clusters), the number of iterations needed to produce an acceptable optimum is too large. In addition, the method is incapable of determining the number of clusters.

The most commonly used sorting method is the K-means algorithm (Hartigan and Wong, 1979). K-means is based on a Ward's within-cluster linkage using a squared Euclidean distance to improve perfor-

mance. The algorithm cycles through the samples, at each step placing one sample into clusters that produce the greatest decreases in the linkage. Computation time is reduced by fixing the centroids at the beginning of each cycle. At the end of each cycle, centroids are recalculated and a new cycle begins. The calculation is completed when no samples can be moved during a complete cycle or when the decrease in similarity falls below an error threshold. This method is limited by the computing resources, which determine the number of cycles that can be performed in a reasonable amount of time. Additional problems are also presented by the inherent instability of the algorithm caused by cumulative similarity oscillations between cycles. This condition may result because of a poorly chosen initial state, or because the centroids are not updated until the end of a cycle. However, with a good initial guess (such as the result from a hierarchical analysis) very good results can be obtained with K-means.

J. Paleont., 67(3), 1993, pp. 486–493

0022-3360/93/0067-0486$03.00

AN EVALUATION OF THE V. J. GUPTA CONODONT PAPERS

GARY D. WEBSTER, CARL B. REXROAD, AND JOHN A. TALENT
Department of Geology, Washington State University, Pullman 99164,
Indiana Geological Survey, 611 North Walnut Grove, Bloomington 47405, and
School of Earth Sciences, Macquarie University, New South Wales 2109, Australia

INTRODUCTION

DISTORTION OF the paleontologic literature in most of the 450 papers bearing V. J. Gupta as author or co-author during the past 30 years has been documented by Agarwal and Singh (1981), Talent (1989a, 1989b, 1989c, 1990a, 1990b, 1990c, in press), Talent et al. (1988, 1989, 1990, 1991), Ahluwalia (1989), Bassi (1989, 1990), Brock et al. (1991), and Radhakrishna (1991). Replies to the charges of fabrication and distortion by Gupta (1989, 1990a, 1990b) were futile attempts to distract the reader, rather than to provide information to refute the charges.

Gupta's fraudulent practices have involved most invertebrate phyla as well as the vertebrates and include fossils of Cambrian to Cenozoic age. Review articles, regional summaries, world paleogeographic reconstructions, etc. are now citing these papers as supporting documents. Only a few of Gupta's papers are based on specimens collected by some of his co-authors and verified by them or independent workers.

Webster (1990) alerted the Pander Society members to the Gupta problem, noting that this involved 60 conodont papers. A request was made to the members for information verifying localities, knowledge of specimens sent to Gupta, etc. There was only one reply, from Budurov and Sudar, verifying some of the papers that they co-authored with Gupta.

We then initiated the following compilation and sent letters to all of Gupta's co-authors of conodont papers requesting verification information. We received replies from 19 of the 42 co-authors. In Pander Society Newsletter #23 Webster et al. (1991) reported that the list of Gupta conodont publications now known is 119. Actually, it is only 118 as we had included one abstract. Most of the increase reflected the inclusion of papers repeating conodont lists or occurrence of specific conodonts and regional compilation and review papers lacking conodont systematics and illustrations.

Study of the Gupta conodont papers shows an interesting insight into the pattern of his continuing fabrications. This pattern is basically followed with both Devonian and Carboniferous conodonts. We will use the Devonian as the example. Italicized numbers refer to the annotated bibliographic listings.

Gupta obtained specimens of the Late Devonian North Evans Limestone fauna of Amsdell Creek, New York, U.S.A., in some manner, perhaps from the collections at Aberystwyth as suggested by Wyatt (1990). He then sent some of the specimens to a conodont worker "claiming" that they were collected in Kashmir. A preliminary note, listing the identified specimens, was published with co-authors, i.e., 1967, *item 98*. Gupta also submitted a preliminary notice of "discovery" of Devonian conodonts from Kashmir to a different journal without co-authors, i.e., 1968, *item 22*. After Gupta received the identification list, he submitted it without co-authors and without reference to the previously co-authored paper for publication, i.e., 1969a, *item 23*. A few years later he used some of the specimens to describe a "newly discovered" Devonian fauna from a second locality, this time in Nepal, i.e., 1975a, *item 35*. He used additional specimens to report the "discovery" of Devonian conodonts from Spiti and correlated the three, reusing some of the same photographs of individual specimens used in *item 35* and claiming them to be from a different locality, i.e., 1975b, *item 36*. Without referencing any of these earlier works he referred to the Lutherwan fauna in a review paper, i.e., 1977, *item 44*. Then with co-authors he used some of these papers as supporting documentation when discussing the geology of a particular area, i.e., 1977, *item 82* and 1981, *item 109*, or discussing other types of Devonian fossils that he claimed were "from" part of the Himalaya, i.e., 1979a, *item 84*. Following this, he used some of the material to salt a sample, and this time with different co-authors, to make a preliminary report of the "discovery of a new Devonian fauna" from another locality in Spiti, i.e., 1982c, *item 3*. This is followed by a paper illustrating the "new find," i.e., 1983, *item 4*. Again with a co-author he cites some of these earlier papers when discussing a Devonian ammonoid fauna from the Himalaya, i.e., 1983, *item 83*. Then as supporting documents and with or without co-authors he cites some of these earlier papers when doing review papers, i.e., 1987a, *item 66*, 1988, *item 75*, 1989, *item 7*, 1991, *item 76*.

The end result of the above shingling and recycling of the Devonian conodonts is 15 papers, totaling 181 pages of fabrication polluting the scientific literature. We suspect that a similar pattern is present in some of his other publications referring to fossils other than conodonts.

Our objective in this report is to evaluate the Gupta conodont papers and inform conodont investigators and the geologic community of the papers that are of a spurious and dubious nature as well as those that have been verified. Hopefully, sufficient citations of this report will continue to be made so that future workers will become aware of the problem and citations of his fabricated work will cease.