NUMERICAL BIAS IN CORRELATION: CARBONATE MODAL DATA, STRATIGRAPHIC THICKNESS DATA, AND GEOCHEMICAL DATA

by

RITA MONAHAN SMITH

DECEMBER 1977

ACKNOWLEDGMENTS

John C. Butler's suggestions and cooperation during the past year are very gratefully acknowledged. His contribution has been considerable and I very much appreciate his willingness to teach me about this subject, which is of special interest to him.

Fred Hilterman's thorough critical review of the manuscript greatly improved it, and I appreciate his contribution.

Dean's willingness(?) to let me vent upon him all the frustrations of the past two years should certainly be acknowledged.

ABSTRACT

Numerical techniques commonly used by geologists to quantitatively ferret out relationships among variables include scattergrams, dendrograms, and principal components analysis. These and other analytical methods are based on a measure of similarity, such as the Pearson product-moment correlation coefficient, r. The r matrix, however, may be subject to numerical bias depending on the type of data from which the matrix is calculated. Consequently, inferences about the geologic significance of the between-variable relationships may be unsound and may reflect nothing more than numerical inevitabilities.

R-mode analyses of a set of carbonate modal data, a set of sedimentary thickness data, and a set of geochemical data show that the interpreter of quantitative analysis must bear in mind not only the constraints of the analytical method but also those of the data treatment.

Among other results, this investigation has determined that closure can induce correlation of the rank type as well as of the product-moment type and that principal components analysis may not be any more efficient a reducer of data than a visual inspection of variances.

Contents

Tables

Tables

Appendix

Figures

INTRODUCTION

As computers have become more accessible to researchers, investigations incorporate which ∧ numerical analysis has more and more frequently appeared in the literature. The trend is undoubtedly irreversible and geology certainly is not unaffected by this technologic development. Quantitative analysis is firmly established as an investigative tool in geologic research, and a substantial percentage of the literature of the discipline is essentially lost to the geologist who lacks understanding of the numerical tools about which he reads or, worse yet, which he uses.

In practice, geologists commonly call upon quantitative techniques to "sort out" relationships among variables. Correlation coefficients (Pearson's product-moment correlation coefficient, r, if the data are normally distributed or a nonparametric coefficient if not) may be computed to serve this purpose. Regardless of the technique, however, the investigator must be able to recognize the absence of association (randomness) among the variables so that meaningful association among variables can be confidently asserted. Meaningful association is conventionally established by testing the coefficients against a parent correlation of zero, that is, against a "null" correlation which represents randomly associated variables. However, for data subject to numerical bias, a parent correlation of zero is not an appropriate measure of randomness. For these data, the conventional interpretation of statistical correlation would yield erroneous conclusions about the variable relationships. It is on this problem that the analysis and discussion in this thesis is brought to bear.

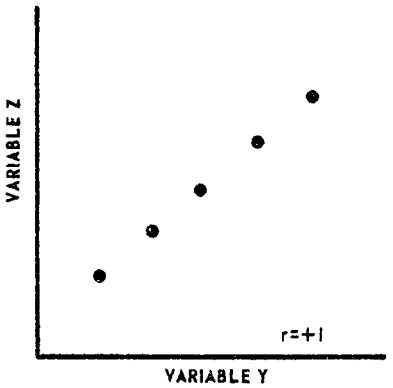
The following section describes techniques which will be used in the analyses in subsequent sections and provides reference material which will be called upon in the discussion of the analyses.

ANALYTICAL METHODS

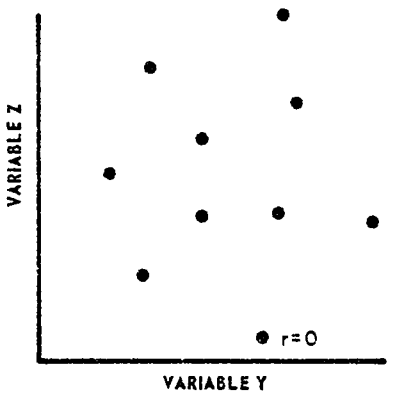
Scattergram, dendrogram, and principal components analysis

By either graphical, intuitive, or numerical methods, many geologic investigators seek to determine relationships between pairs of variables. The binary scattergram is probably the most common and simplest method of illustrating the degree of linear association between a pair of variables because the joint behavior of the two variables is assessable by simple graphical interpretation. The scatter of points in figure 1 shows a perfect positive linear association (fig. 1A), a complete lack of linear association (randomness)(fig. 1B), and a perfect negative linear association (fig.1C). More realistic patterns of joint variation deviate from the perfect relationships shown in figure 1, but it is these standards of reference against which real data are compared. By simply plotting points on a scattergram, then, the investigator may visually evaluate the strength of pairwise association between variables.

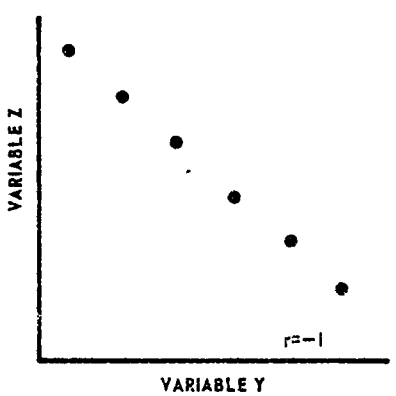
A multivariate counterpart of the binary scattergram is the dendrogram, produced by cluster analysis. A dendrogram is a "tree diagram" showing the mutual relationships among a given set of variables. The variables most highly intercorrelated are clustered together in a manner such that the dendrogram has a hierarchical configuration in which each level reflects the degree of intragroup homogeneity. Mutually related groups are adjacent, but the groups are mutually exclusive in that any variable can belong to only one group. McCammon (1969) stated, however, that it is possible to examine the data in a rearranged form in which



A. POSITIVE ASSOCIATION



B. RANDOM ASSOCIATION (UNCORRELATED)



C. NEGATIVE ASSOCIATION

FIGURE 1. – – PAIRWISE ASSOCIATION PATTERNS.

variables that could interchange groups can be identified.

Davis (1973) commented that cluster analysis is not a statistically rigorous method of examing multivariate behavior. Rather, he stated, "it belongs to that category of techniques...in which utility is judged by performance and not by theoretical considerations" (p. 501). It tends to be used insofar as its results corroborate the investigator's intuition about the variables.

Principal components analysis (PCA) also falls into the category of techniques in which utility is judged by performance. PCA is not, strictly speaking, a statistical procedure. It is a mathematical manipulation designed to make apparent the "redundancy" in a set of variables. The manipulation transforms the data to a new coordinate system defined to produce uncorrelated variables. As a multivariate technique, it could perhaps serve as a more quantitative discriminator--relative to cluster analysis---of the interrelationships among variables.

Measure of similarity

The binary scattergram, the dendrogram,and PCA are all techniques of R-mode analysis as long as the objective is to seek interrelationships among the variables. A measure of similarity, that is, a measure of pairwise variation among the variables, is the basis for analysis. The measure of pairwise variation that is commonly chosen is a matrix of correlation coefficients, either the Pearson product-moment correlation coefficient (r) if the data are normally distributed or a rank correlation coefficient if not. The product-moment correlation coefficient is a measure of linear association and is defined as the ratio of the covariance (cov) of two variables (e.g., y and z) to the

product of their standard deviations (s):

$$r_{yz} = \frac{cov_{yz}}{s_y s_z} \qquad (1)$$

This measure is a convenient one because, being unitless, r's are easily compared even though the variables may have been measured in different units. Moreover, r is a "standardized" measure of similarity; that is, each variable is weighted equally. The value of r can range from +1 to -1(because covariance yz may equal but not exceed the product of the standard deviations of its variables). Whereas r = +1 indicates a perfect direct relationship and r =-1 indicates a perfect inverse relationship, r = 0 indicates complete linear randomness. These interpretations of r, especially in combination with a binary scattergram (fig. 1), conform well with what an investigator would intuitively conclude about the pairwise association patterns.

Intuitive conclusions may be statistically corroborated by testing the observed correlation coefficients against the null hypothesis that the correlation in the target population is zero. This criterion of randomness is, in fact, the assumption underlying most elementary correlation analysis; one of its major advantages is the compatibility between the numerical and graphical representations of association. If an observed correlation differs with statistical significance from zero, the variables are meaningfully associated. The test for significance of r is made using Student's t (Neter and Wasserman, 1974); $t_{calculated}$ is the test statistic:

$$H_o: \quad \rho = 0$$

$$H_a: \quad \rho \neq 0$$

$$t_{calculated} = \left| r_{observed} \left[(n-2)/(1-r^2_{observed}) \right]^{\frac{1}{2}} \right|$$

where n is the number of samples. If $t_{calculated}$ is greater than $t_{tabulated}$ (at the given significance level and for the appropriate degrees of freedom), then the null hypothesis H_o must be rejected, the implication of which is that no reason exists to assume the variables are not meaningfully associated.

DATA TREATMENT

For some types of data, the "intuitive demand" (Chayes, 1971) that zero correlation be the measure of randomness is ill-conceived because the matrix of correlation coefficients is vulnerable to substantial numerical bias. Bias alters the measure of randomness so that meaningful association among variables is obscured. Consequently, inferences about the geologic significance of the variable relationships may be unsound and may reflect nothing more than numerical inevitabilities. Data which are subject to induced bias are those in percentage and ratio form.

Percentage (closed) form

Consider that on each of n samples m variables are measured and these data are arranged in matrix form such that each row represents one sample and each column represents one variable. From this matrix, a deviation matrix can be formed by subtracting the variable mean from each of the measured values of that variable. The data are "open" if no constraints are imposed on the sum of the rows of the data matrix or of the deviation matrix. The data are "closed" if the sum of each row of the data matrix is constrained to be a constant common to all rows. The constant sum is 100 in the case of percentage data. Moreover, for closed data the sum of each row in the deviation matrix is constrained to be zero. Therefore $\wedge$ the sum of each row in the closed variance-covariance matrix must be zero (no such restriction applies to the sums of the rows of the open variance-covariance matrix)(Chayes, 1971, p. 37). This implies

that the sum of all covariances for variable x is negative and has the same magnitude as the variance of x. Each variable, therefore, must be negatively correlated with at least one other variable. Generally, strong negative correlation bias is induced between variables having relatively large variance, and positive correlation may be induced between variables whose open variances are relatively small (Chayes, 1971, p. 39-40).

Essentially, then, the effect of closure (percentage formation) is that the potential independence of variances and covariances (that is, the potential for covariances to be zero), which is the basis of correlation theory, is destroyed; bias is introduced such that the closed data are statistically correlated even though the open data may not have been.

For two- and three-variable systems, the correlation bias due to percentage formation can be exactly predicted. The binary case is trivial; the induced correlation is -1, a perfect inverse linear association, because the relative magnitude of one variable must decrease as the other increases. In the ternary case, all three correlation coefficients are fixed and predictable from the (closed) variances (var) of the variables (Chayes, 1971, p. 42):

$$r_{yz} = \frac{var_x - (var_y + var_z)}{2s_y s_z} \qquad (2)$$

where s is the standard deviation. (For the other coefficients in the ternary system, r_{xz} and r_{xy}, the subscripts in equation 2 can be rotated.) It is apparent that r_{yz} can be positive only if var_x is greater than $(var_y + var_z)$; the implication is that positive correlation in a ternary

system would not be a common occurrence. Note that the variance of the third variable (the one excluded from r's subscripts (equation 2)) controls the sign of r.

For multivariate systems the correlations due to closure cannot be exactly predicted. Chayes (1960, 1971) has shown, however, that for systems having four or more variables, bias is not negligible. Moreover, one cannot expect to eliminate the closure effect by making the number of variables (m) very large. Only in the unlikely case in which all open variables are randomly distributed and have the same mean and variance can the effects of closure be reduced by increasing the number of variables. Chayes (1971, p. 40) has shown that, for m varibles so defined, the closed correlation coefficient to be expected between any two is:

$$r = -\frac{1}{(m-1)} \qquad (3)$$

For example, suppose that

m = 3; then r =-.5 in the absence of a significant departure from randomness;

if m = 5; then r =-.25

m =10; r =-.11

m =15; r =-.07

This extreme case, however, is hardly realistic among geologic variables.

To illustrate the effect of closure, a data set of 75 samples, each sample having 5 normally distributed uncorrelated variables, was simulated. The summary statistics for the open and closed data are given in table 1. Figure 2 shows examples of binary scattergrams of the open (figs. 2A,B) and the closed (figs. 2C,D) data. It is obvious that closure has induced linear association and has obscured the picture of randomness, thereby obscuring meaningful association among the variables.

Table 1.--Statistics describing simulated set of 5 normally distributed random variables measured on 75 observations

Stratigraphic thickness measurements (open set)

Summary statistics

Variable name	A	B	C	D	E	Total thickness
Mean	147.8	149.8	100.3	48.84	298.6	745.3
Variance	222.9	24.46	94.34	118.0	5606.	5915.
Standard deviation	14.93	4.946	9.713	10.86	74.88	76.91
Coefficient of variation	.10099	.03302	.09683	.22240	.25079	.10319
Maximum	176.6	166.2	122.1	77.60	485.1	942.3
Minimum	102.7	137.7	80.86	21.30	118.8	573.5
Range	73.97	28.53	41.20	56.30	366.3	368.8
Percent variance	1.8607	.20415	.78736	.98486	46.794	49.369

Variance - covariance matrix

	A	B	C	D	E	Total thickness
A	222.93	1.0384	4.7960	6.9150	-145.97	89.713
B		24.460	2.6881	6.9054	-52.066	-16.976
C			94.335	6.3200	-2.2914	105.84
D				118.00	96.070	234.21
E					5606.5	5502.2
Total thickness						5915.0

Correlation - coefficient matrix

	A	B	C	D	E	Total thickness
A	1.0000	.01406	.03307	.04264	-.13056	.07813
B		1.00000	.05596	.12854	-.14060	-.04463
C			1.00000	.05990	-.00315	.14169
D				1.00000	.11812	.28034
E					1.00000	.95547
Total thickness						1.00000

Percent of total section (closed set)

Summary statistics

Variable name	A	B	C	D	E
Mean	20.03	20.32	13.58	6.583	39.49
Variance	7.527	5.267	3.185	2.107	39.96
Standard deviation	2.743	2.295	1.785	1.452	6.321
Coefficient of variation	.13698	.11296	.13141	.22052	.16006
Maximum	25.37	27.09	19.41	9.942	54.32
Minimum	13.39	15.74	9.623	3.648	20.71
Range	11.98	11.35	9.786	6.894	33.61
Percent variance	12.968	9.0738	5.4879	3.6309	68.840

(continued)

Table 1.--(continued)

Variance - covariance matrix

	A	B	C	D	E
A	7.5267	4.1102	2.2314	.59584	-14.464
B		5.2666	2.7198	.85132	-12.947
C			3.1852	.42638	- 8.5625
D				2.1074	- 3.9808
E					39.956

Correlation - coefficient matrix

	A	B	C	D	E
A	1.0000.	.65283	.45572	.14961	-.83405
B		1.00000	.66405	.25554	-.89255
C			1.00000	.16457	-.75900
D				1.00000	-.43382
E					1.00000

Ratios of closed variables[1]; denominator has large coefficient of variation

---

Summary statistics

Variable name	A/E	B/E	C/E	D/E
Mean	.53173	.53893	.36016	.17447
Variance	.02660	.02716	.02187	.00385
Standard deviation	.16309	.16481	.11344	.06203
Coefficient of variation	.30671	.30581	.31497	.35552
Maximum	1.19175	1.30800	.93695	.39082
Minimum	.26007	.30919	.17806	.07153
Range	.93168	.99880	.75889	.31929
Percent variance	37.74040	38.54042	18.25990	5.45928

Variance - covariance matrix

	A/E	B/E	C/E	D/E
A/E	.02660	.02529	.01648	.00753
B/E		.02716	.01765	.00803
C/E			.01287	.00524
D/E				.00385

Correlation - coefficient matrix

	A/E	B/E	C/E	D/E
A/E	1.00000	.94100	.89083	.74391
B/E		1.00000	.94399	.78514
C/E			1.00000	.74438
D/E				1.00000

[1] (equal to ratios of open variables)

(continued)

Table 1.--(concluded)

Ratios of closed variables; denominator has small coefficient of variation

Summary statistics

Variable name	A/B	C/B	D/B	E/B
Mean	.98806	.67027	.32614	1.99773
Variance	.01105	.00451	.00517	.26176
Standard deviation	.10514	.06715	.07190	.51163
Coefficient of variation	.10641	.10019	.22045	.25610
Maximum	1.19886	.83700	.50812	3.23421
Minimum	.70339	.53664	.14025	.76453
Range	.49547	.30036	.36787	2.46968
Percent variance	3.91285	1.59624	1.82995	92.66096

Variance - covariance matrix

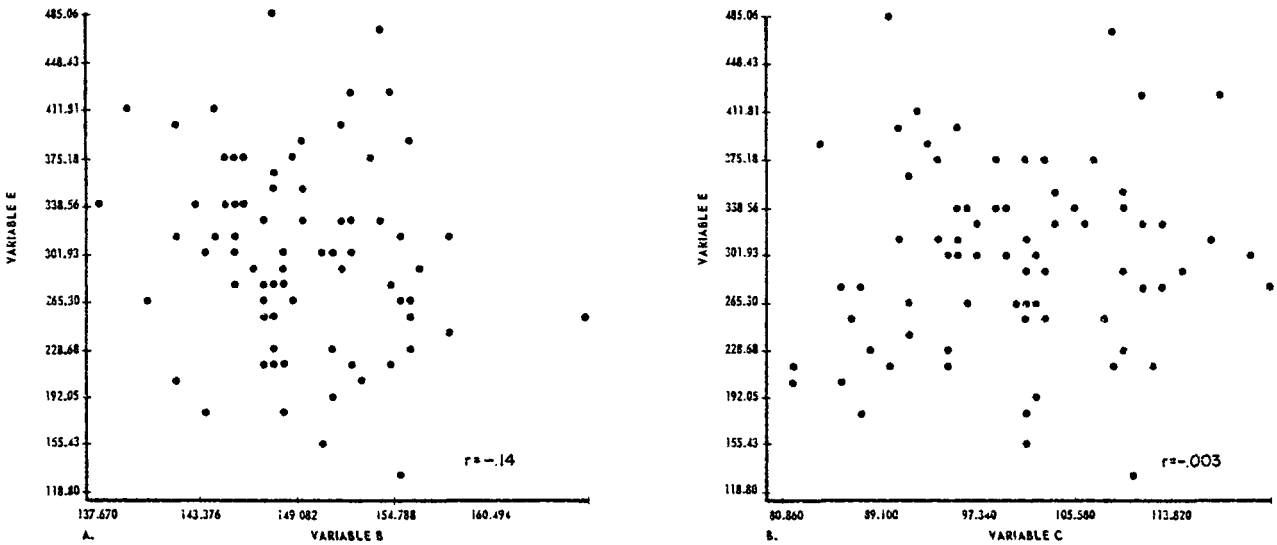
	A/B	C/B	D/B	E/B
A/B	.01105	.00077	.00027	-.00207
C/B		.00451	.00025	.00243
D/B			.00517	.00553
E/B				.26176

Correlation - coefficient matrix

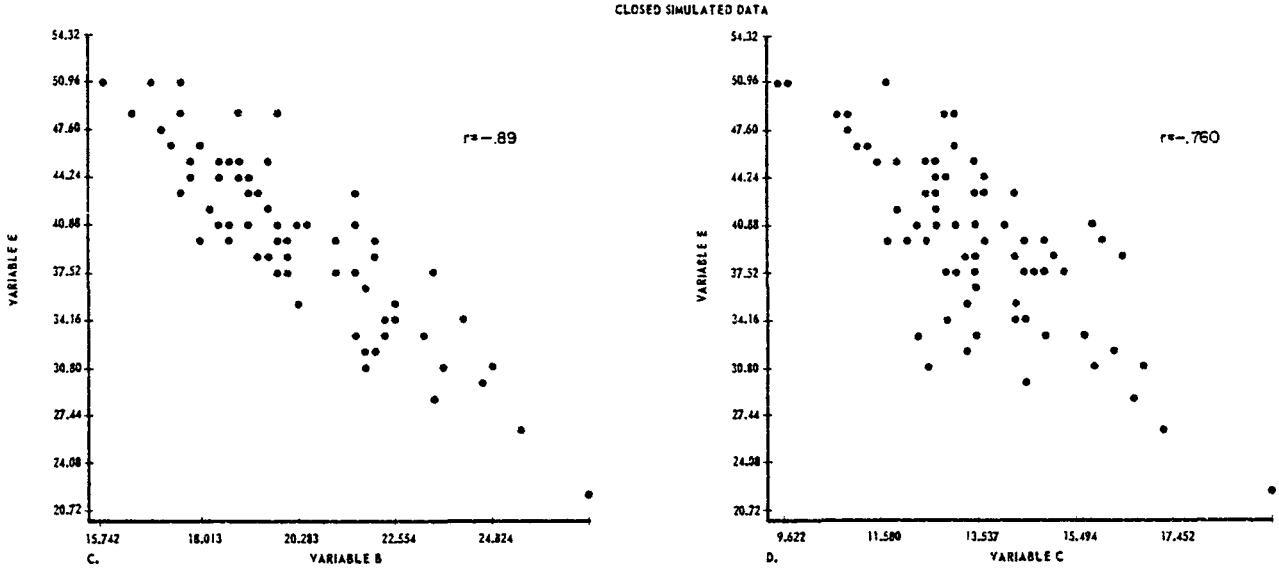
	A/B	C/B	D/B	E/B
A/B	1.00000	.10923	.03593	-.03852
C/B		1.00000	.05233	.07081
D/B			1.00000	.15040
E/B				1.00000

FIGURE 2. – – BINARY SCATTERGRAMS OF OPEN AND CLOSED SIMULATED DATA.

OPEN SIMULATED DATA



CLOSED SIMULATED DATA



Ratios having common terms

As will be shown, the closure effect can be eliminated by converting percentage data to ratios. However, the formation of ratios from percentages essentially replaces the constraints of closure by those inherent in ratio formation.

Suppose that X is a vector of random variables $[x_1, x_2, \ldots x_m]$ and that P is a vector of percentages such that $P_k = x_k/rs$ where rs (row sum) $= \sum_{k=1}^{m} x_k$ and $1 \leq k \leq m$. Q_i is defined as the ratio of any two P_k's:

$$Q_i = \frac{P_1}{P_2} = \frac{\frac{x_1}{rs}}{\frac{x_2}{rs}} = \frac{x_1}{x_2}$$

Similarly, let Q_j be defined as:

$$Q_j = \frac{P_3}{P_2} = \frac{\frac{x_3}{rs}}{\frac{x_2}{rs}} = \frac{x_3}{x_2}$$

Therefore ratios of percentages (recall that percentages are subject to the closure effect) "reduce" to ratios of nonpercentage variables. Because it evidently does not matter whether the variables from which ratios are formed are percentages or not, a discussion of the correlation constraints of ratio formation can proceed assuming the parent variables (e.g., x_1, x_2, etc.) are not percentages and that they are uncorrelated. As will be shown, correlation between simple ratios having common terms is measuring bias as well as the degree of association between the variables. (No correlation will be induced between two simple ratios which do not have common terms.)

Furthermore, it will be shown that various (depending on the ratio form) nonzero null values can be derived to test the significance of correlations between ratios.

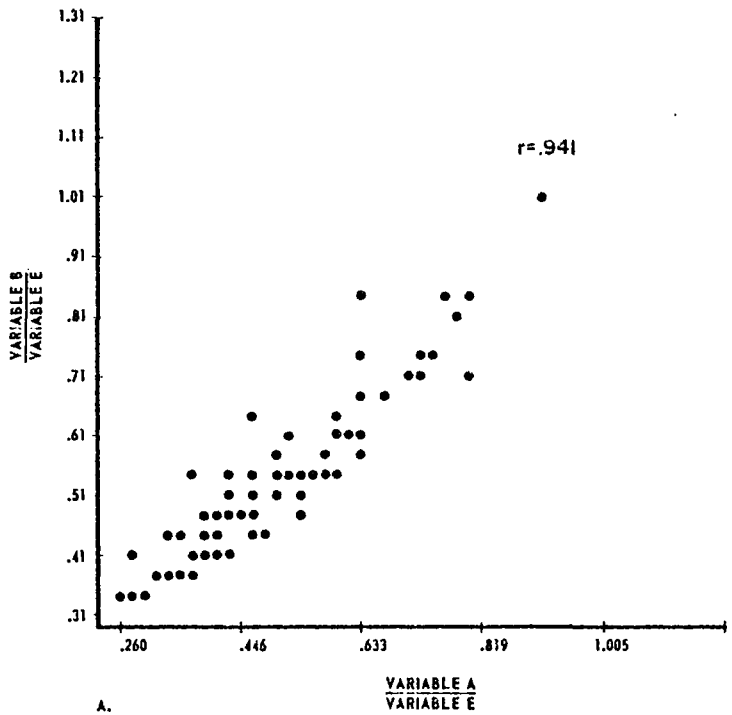
Pearson (1896) showed that, even when variables are uncorrelated, "spurious" correlation between ratios formed from the variables is a function of the means, variances, and covariances of the variables forming the ratios. Pearson's general formula for the correlation (r_{ij}) between ratios Q_i and Q_j (as defined in table 2) is given in equation 1 of table 2. This equation has been used by Chayes (1949, 1971) to derive approximations (equations 2-6, table 2) for the correlation to be expected between various common forms of ratios constructed from uncorrelated (or correlated, as in the case of percentages) variables.

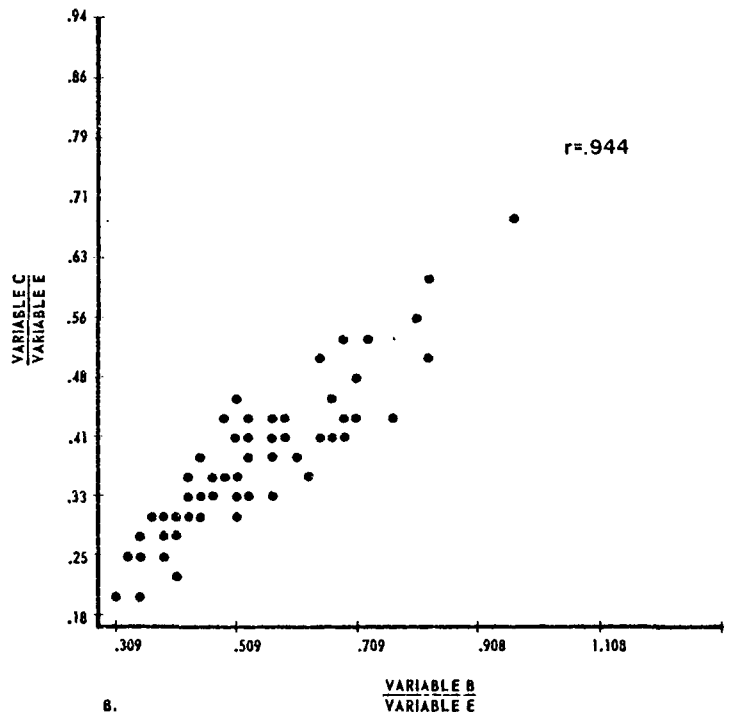
To illustrate the problem of ratio correlation, the simulated data described in table 1 is plotted in figure 3. The ratio form corresponding to equation 4 (table 2) was chosen for figure 3 because ratios with common denominators are often used as a scaling mechanism. The denominator, variable E, is the variable having the largest coefficient of variation, C, where $C_E = s_E/\bar{x}_E$, where s is the standard deviation, and where $\bar{x}$ is the mean of the variable. The figures (3A,B) show that the ratio formation has induced remarkably high correlations. It can be shown (Chayes, 1949) that, when the variable having the largest coefficient of variation is in the denominator for this ratio form, induced correlation can be no less than $r_{ij} = .5$, and this minimum applies only in the special case when $C_1 = C_2 = C_3$ (table 2, equation 4). As the data departs from this unlikely case, induced correlation rises rapidly. For example,

Table 2.--"Spurious" correlation among ratios

Equation No.	Pearson's general formula for "spurious" correlation
	Suppose X is a vector of k random variables $(X_1, X_2 \ldots X_m)$. Q_i is a ratio formed from any two X_k's and Q_j also is a ratio formed from two X_k's. If $Q_i = \frac{X_1}{X_2}$, $Q_j = \frac{X_3}{X_4}$, $C_k = \frac{S_k}{\bar{X}_k}$ (where C is the coef. of variation, S is the standard deviation, and $\bar{X}$ is the mean), and $r_{k'}$ is the observed correlation coef. between X_k and $X_{k'}$,
1	then $r_{ij} = \frac{r_{13}C_1C_3 - r_{14}C_1C_4 - r_{23}C_2C_3 + r_{24}C_2C_4}{(C_1^2 + C_2^2 - 2C, C_2r_{12})^{1/2} (C_3^2 + C_4^2 - 2C_3C_4r_{34})^{1/2}}$
	Approximate null correlations between ratios having common terms
	Suppose X is a vector of k random variables $(X_1, X_2 \ldots X_m)$. Q_i is a ratio formed from any two X_k's and Q_j is a ratio formed from two X_k's, one of which is common to Q_i or one of which is common and one of which is a constant (e.g., 1); $C_k = \frac{S_k}{\bar{X}_k}$ as defined for equation 1.
2	If $Q_i = \frac{X_1}{X_2}$ and $Q_j = X_1$, then $r_{ij} = \frac{C_1}{(C_1^2 + C_2^2)^{1/2}}$
3	$Q_i = \frac{X_1}{X_2}$ and $Q_j = X_2$, $r_{ij} = \frac{-C_2}{(C_1^2 + C_2^2)^{1/2}}$
4	$Q_i = \frac{X_1}{X_2}$ and $Q_j = \frac{X_3}{X_2}$ $r_{ij} = \frac{C_2^2}{(C_1^2 + C_2^2)^{1/2} (C_3^2 + C_2^2)^{1/2}}$
5	$Q_i = \frac{X_2}{X_1}$ and $Q_j = \frac{X_2}{X_3}$ $r_{ij} = \frac{C_2^2}{(C_2^2 + C_1^2)^{1/2} (C_2^2 + C_3^2)^{1/2}}$
6	$Q_i = \frac{X_1}{X_2}$ and $Q_j = \frac{X_2}{X_3}$ $r_{ij} = \frac{-C_2^2}{(C_1^2 + C_2^2)^{1/2} (C_2^2 + C_3^2)^{1/2}}$

FIGURE 3. -- BINARY SCATTERGRAMS OF SIMULATED DATA IN RATIO FORM WHERE DENOMINATOR HAS LARGE COEFFICIENT OF VARIATION.



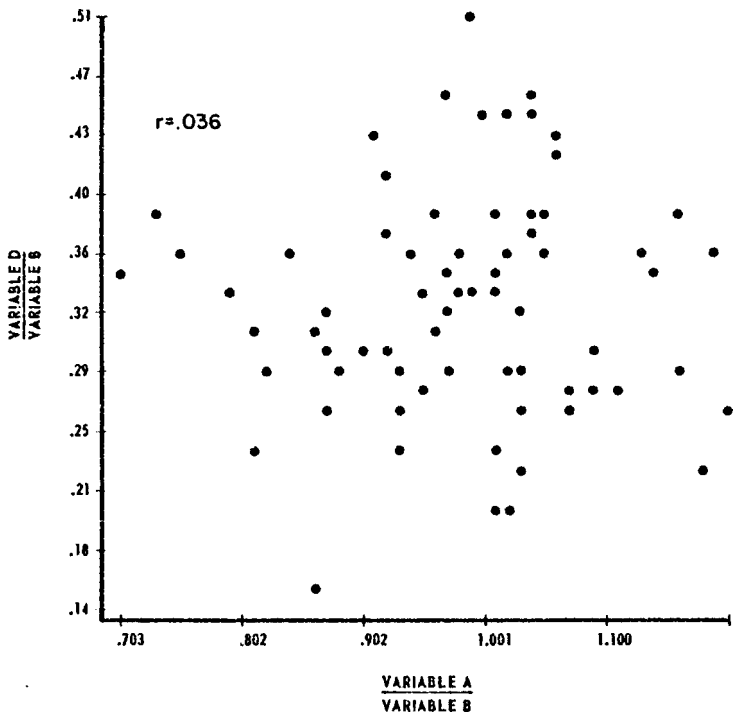


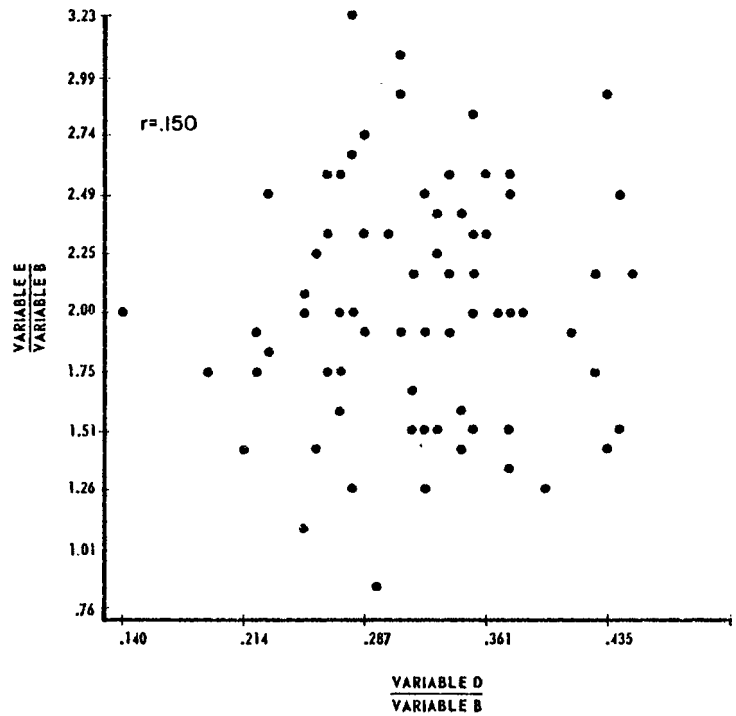
if C_2 is twice as large as C_1 and C_3, $r_{ij} = .67$; if C_2 is three times as large as C_1 and $1\frac{1}{2}$ times as large as C_3, $r_{ij} = .79$. If, on the other hand, the variable in the denominator of this ratio form has a very small C (e.g., variable B, table 1), very little correlation is induced (fig. 4A,B).

It is obvious that data treatment (percentage or ratio formation) may jeopardize an investigator's ability to recognize lack of linear association among variables, but he must nevertheless avoid attributing significance to bias (correlations) generated by the data treatment. Therefore, rather than test observed correlations against a null value of zero, Chayes (1971) has suggested that the "spurious" correlation itself (the correlation induced among the parent variables defined to have covariances equal to zero) serve as the measure of randomness, the null value. If an observed correlation differs significantly from the spurious correlation, the investigator may conclude that, at the given significance level, the variables are meaningfully associated. The point, of course, is that the investigator must know the correlation to be expected in the absence of a significant departure from randomness.

The following sections report R-mode analyses of a set of carbonate modal data, a set of sedimentary thickness data, and a set of geochemical data. The analytical (statistical) techniques and the data characteristics vary from set to set, but the common element is that the constraints of the analytical method cannot be separated from constraints of the data treatment; both must be brought to bear on the interpretation of an analysis.

FIGURE 4. -- BINARY SCATTERGRAMS OF SIMULATED DATA IN RATIO FORM WHERE DENOMINATOR HAS SMALL COEFFICIENT OF VARIATION.





CARBONATE MODAL DATA

Purdy (1960) conducted an investigation "to quantitatively delineate calcium carbonate facies on the northwestern part of the Great Bahama Bank" (1960, p. 1). His work is considered a classic study in quantitative analysis of geologic data. In fact, according to Davis (1973), the use of Q-mode factor analysis was introduced into geology by Imbrie and Purdy (1962) using the data of Purdy (1960). Because the dissertation is so widely cited (e.g., McCammon, 1969; Parks, 1966; Koch and Link, 1971; Krumbein and Graybill, 1965) and because it raises some interesting questions, a reexamination of Purdy's analysis follows.

This section describes Purdy's investigation and suggests that some assumptions under which he worked were not rigorously justifiable. This thesis attempts to identify incorrect assumptions and to insure that the analytical methods are appropriate to the data. In spite of the analytical differences, the results herein are generally comparable to Purdy's. Interpretive differences, however, are introduced as a result of the analysis in this section of the effect of closure on the data.

Purdy's investigation

Purdy collected 218 sediment samples, from each of which a representative subsample was selected and thin-sectioned. For each of the 218 subsamples, the constituent particle composition (the relative abundance of 16 different grain types in a sample) was determined by point-count analysis

of the sample fraction coarser than 1/8 mm. Another representative subsample was selected to measure the weight percentage of the sample fraction finer than 1/8 mm (the boundary between the fine-sand and very-fine-sand sizes of the Wentworth scale). This weight percentage was intended to be a textural indicator which would reflect (although, as Purdy conceded, crudely) the relative intensity of current action in different areas on the bank.

Purdy's data formed a matrix of 17 variables measured on 218 samples, but not all this information was included in his statistical analysis. Four variables were excluded because they were not "quantitatively important constituents"; 15 samples were eliminated because of thin-section analytical error due to relatively large grain size. Therefore, Purdy's working matrix included 203 samples and 13 variables. His first objective was to determine which variables tend to react similarly to various environmental conditions. Such variables form what Purdy called a "reaction group". He stated that "the distribution of the various reaction groups constitutes a sedimentary facies", which he defined as "areally segregated parts of differing nature belonging to any genetically related body of sedimentary deposits" (p. 94).

To resolve his data into reaction groups, Purdy used the technique of cluster analysis and chose the correlation coefficient r to be the measure of similarity between variables. He stated that the significance of the resulting dendrogram "is that it documents the extent to which Bahamian constituents tend to occur or react together in various unspecified environments" (1960, p. 87). His analysis discriminated four reaction groups.

The use of Pearson's r as a measure of pairwise variation between variables requires that each variable is normally distributed and that both variables together are bivariately normally distributed. Purdy stated that "justification for the use of this correlation coefficient [in his study] is found in the central-limit theorem which states in part that '. . .as sample size increases, sample means tend to be distributed normally even is the parent population is anormal' (Snedecor, 1965, p. 71). If the distribution is not markedly skewed, the approximation to the normal distribution will usually be sufficiently good if the sample size exceeds 30 (Cramer, 1955, p. 184). . .Sample size in the present study is 203; moreover the estimated volume abundance of grain types in each thin section is based on at least 500 point-counts. Therefore it seemed fairly safe to compute product-moment correlation coefficients" (Purdy, 1960, p. 87).

But perhaps it is not safe. According to the central-limit theorem, the distribution of the means of all possible sample populations of size n of a nonnormal target population will approach normality if the number of samples (n) per sample population is sufficiently large. Even 30 samples is "sufficiently large" if the target population is "not badly 'nonnormal' " (Kolstoe, 1973, p. 145). This is not the same as saying that the distribution of any one sample population will approach normality if n is sufficiently large, which is the interpretation of the central-limit theorem by which Purdy justified the use of r. Moreover, Purdy's data are markedly skewed, and nonnormality precludes the use of r. Histograms of all 17 variables show that none are normally distributed; all are extremely skewed toward zero occurrence. The chi-square goodness-of-fit test (Davis, 1973, p. 116-122) for each variable corroborates what the

histograms suggest. Therefore a measure of similarity which assumes a normal distribution is dubiously applicable to Purdy's data.

Rank correlation coefficients and reaction groups

Nonparametric statistics are applicable regardless of the parameters of the target population from which the sample is drawn (Till, 1974); that is, their use is not contingent upon any assumptions about the underlying distribution of the data. Rank correlation coefficients, such as Spearman's r_s or Kendall's τ (Demirmen, 1976) are nonparametric "order" statistics computed from ranks rather than from absolute scores (as Pearson's r is computed). They measure the degree of agreement between the ranks. Tests for variable association which are based on these coefficients are nonparametric. In general, numerical values of Spearman's and Kendall's coefficients computed from the same data set are not identical because the exact form of association measured by the two coefficients is different. Generally the absolute value of Spearman's coefficient exceeds that of Kendall's, but the product-moment correlation between the two coefficients (for the same data set) in the null situation is high, approaching unity for large sample size (Demirmen, 1976, p. 223).

Because Purdy's data is not normally distributed, the rank correlation coefficients would be more appropriate measures of similarity for the data than the product-moment coefficient. Both rank coefficients were computed (using RANK from Demirmen, 1976) for Purdy's 203 x 13 matrix, and the results are given in table 3a. As expected, in every case but one (that of the correlation between grapestone and oolites), the absolute value of Spearman's coefficient exceeds that of Kendall's. Table 3b shows that both rank coefficients are very different from the product-moment coefficient for the same variables.

Table 3a.--Rank-correlation-coefficient matrix for observed carbonate modal data (Spearman's coefficient is in upper-right half of matrix; Kendall's coefficient is in lower-left half)

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains	Sample fraction < 1/8 mm
Coralline algae	1.000	.330	.134	.232	.531	.196	-.206	.013	.132	.285	-.186	.331	-.111
Halimeda	.258	1.000	.527	.533	.444	.703	.275	.369	-.049	.255	-.486	.111	.304
Peneroplidae	.106	.382	1.000	.657	.253	.660	.504	.560	-.177	.244	-.685	-.053	.595
Other forams	.183	.397	.497	1.000	.177	.614	.399	.506	.054	.376	-.727	.282	.594
Corals	.483	.356	.202	.140	1.000	.307	-.126	.102	.050	.246	-.213	.179	-.143
Molluscs	.151	.527	.491	.456	.243	1.000	.347	.392	.005	.374	-.753	.109	.378
Fecal pellets	-.164	.188	.358	.283	-.105	.239	1.000	.767	-.329	-.056	-.416	-.246	.770
Mud aggregates	.013	.260	.411	.365	.081	.277	.578	1.000	-.264	.100	-.451	-.140	.677
Grapestone	.102	-.043	-.125	.038	.036	.001	-.229	-.187	1.000	.495	-.122	.811	-.328
Organic aggregates	.238	.187	.181	.280	.202	.271	-.047	.071	.359	1.000	-.491	.470	-.025
Oolites	-.148	-.351	-.516	-.553	-.168	-.578	-.299	-.326	-.125	-.365	1.000	-.288	-.477
Cryptocrystalline grains	.257	.074	-.031	.197	.139	.077	-.176	-.101	.622	.346	-.238	1.000	-.161
Sample fraction < 1/8 mm	-.085	.225	.449	.431	-.112	.284	.571	.501	-.226	-.021	-.351	-.106	1.000

Table 3b.--Pearson-product-moment correlation-coefficient matrix for observed carbonate modal data (from Purdy, 1960)

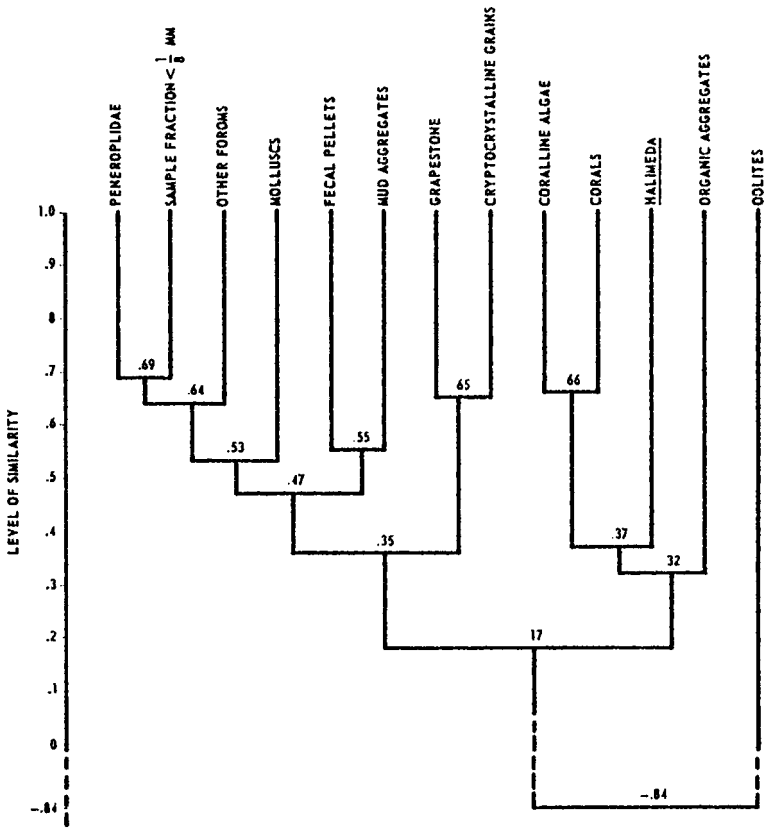
	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains	Sample fraction < 1/8 mm
Coralline algae	1.000	.323	-.001	.139	.662	.135	-.195	-.034	-.072	.391	-.199	.167	-.125
Halimeda		1.000	.141	.275	.349	.411	.012	.188	-.201	.152	-.351	-.051	.177
Peneroplidae			1.000	.612	.006	.450	.308	.316	-.264	.068	-.399	-.184	.687
Other forams				1.000	.012	.453	.129	.302	-.116	.230	-.520	.137	.558
Corals					1.000	.327	-.170	.045	-.106	.221	-.206	.057	-.149
Molluscs						1.000	.123	.221	-.138	.136	-.541	.001	.464
Fecal pellets							1.000	.546	-.401	-.206	-.372	-.449	.657
Mud aggregates								1.000	-.321	.017	-.410	-.261	.586
Grapestone									1.000	.235	-.307	.645	-.410
Organic aggregates										1.000	-.337	.332	-.098
Oolites											1.000	-.423	-.435
Cryptocrystalline grains												1.000	-.371
Sample fraction < 1/8mm													1.000

One might expect that cluster analysis, as it is based on the measure of similarity, would reflect the difference between the two types of coefficients.

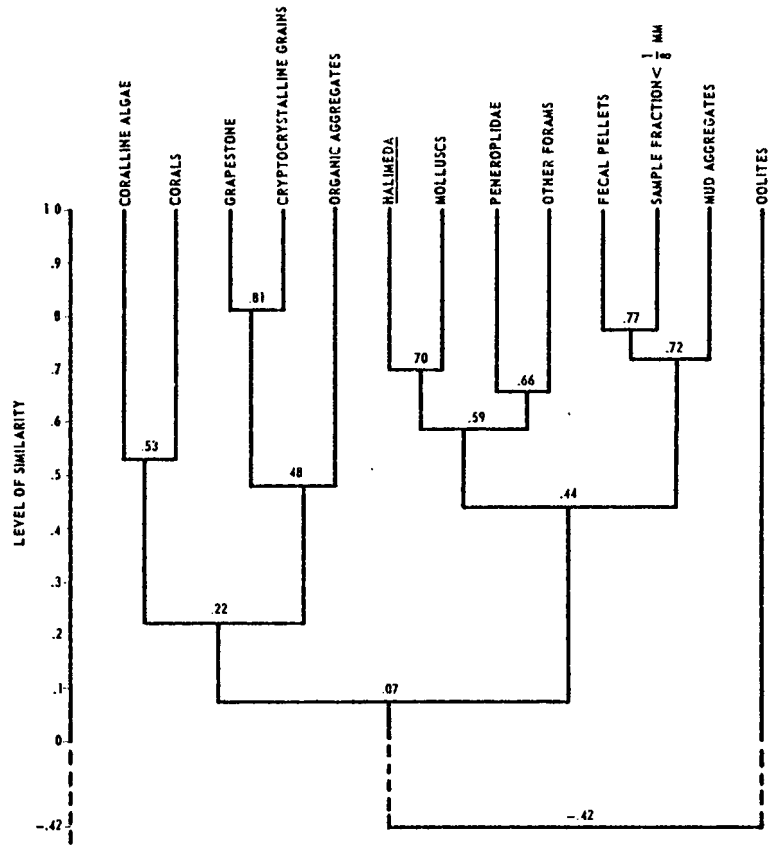
To bring information to bear on this speculation, the matrix of Spearman's rank correlation coefficients was clustered using the program CLUSTER (Davis, 1973). Spearman's coefficient was chosen over Kendall's because, according to Demirmen (1976), Spearman's coefficient is "in effect a product-moment correlation coefficient obtained by treating the ranks as though they were actual scores. Thus this coefficient in a sense measures the degree of linear relationship between the ranks of two variables . . . high values of (Spearman's coefficient) indicate that the basic form of the relation between two variables is monotone, i.e., an increase in one variable is accompanied by an increase or decrease in the other variable, although not necessarily in a linear manner" (p. 223). (A cluster made using Kendall's coefficient for Purdy's data, however, was not noticeably different from that made using Spearman's coefficient, except that the levels of similarity were lower for the former.)

Results of the rank clustering are shown in figure 5B and may be compared with Purdy's dendrogram, in figure 5A. Even the subjective nature of dendrogram interpretation could not interfere with the very obvious similarities between the two diagrams. As the summary in table 4 shows, for all practical purposes the reaction groups of Purdy can be reclaimed by clustering the rank correlation coefficients of the data.

FIGURE 5. – – DENDROGRAMS OF OBSERVED CARBONATE DATA.



A. DENDROGRAM FOR r AS MEASURE OF SIMILARITY
(FROM PURDY, 1960)



B. DENDROGRAM FOR RANK CORRELATION COEFFICIENT
AS MEASURE OF SIMILARITY

Table 4.--Reaction groups for observed carbonate modal data

Using Pearson's product-moment coefficient		Using Spearman's rank coefficient	
Group II:	grapestone and cryptocrystalline grains	--	grapestone and cryptocrystalline grains
			organic aggregates
III:	coralline algae and corals	--	coralline algae and corals
	Halimeda		
	organic aggregates		
IV:	oolites	--	oolites
I:	Peneroplidae and sample fraction < 1/8mm	--	Halimeda, molluscs, and Peneroplidae
	other forams		other forams
	molluscs		sample fraction < 1/8 mm
	fecal pellets		fecal pellets
	mud aggregates		mud aggregates

It should be noted, however, that the reaction groups which Purdy explicitly defined (1960, p. 94 ff) as in table 4 were assigned in a manner inconsistent with the conventional interpretation of cluster diagrams. Dendrograms yield groupings, although subjectively, according to a chosen cut-off level of similarity. No rigorous statistical method dictates what level an investigator must choose, but it is conventional that only one level be used to establish the groupings. The intent of this convention is to minimize the introduction of the investigator's personal bias into the interpretation of the dendrogram.

Inspection of figure 5A shows that Purdy's reaction groups could not be as he defined them if only one level had been used for the interpretation. Groups I and II, as defined, require a cut-off level of similarity of $.35 < r < .47$; group III, as defined, dictates a cut-off of $r < .32$. Had $r = .36$, for example, been the discriminating level of similarity across the dendrogram, organic aggregates, as well as oolites, would have been a one-variable group.

Closure and reaction groups

Purdy's data are modal (that is, they are percentages) and are therefore subject to the constant-sum restraint. Because of closure, negative correlation between some of the variables has inevitably been induced. One would reasonably expect this bias to be reflected in a cluster analysis based on a measure of correlation. To assess the closure effect, one must know the correlation to be expected in the absence of a significant departure from randomness. This is the null correlation against which an observed correlation is tested for significance. In the following analysis, simulation is used to establish the null correlations.

Actually only 12 of Purdy's 13 "quantitatively important" variables are modal; weight percentage of the sample fraction finer than 1/8 mm is a textural indicator only and is not a variable of constituent particle composition. The 12 quantitatively important modal variables were recomputed to 100 percent, and the summary statistics describing the recomputed data are given in table 5. These statistics are used in the simulation to assign values to the parameters (means and variances) of a hypothetical open matrix. The hypothetical open matrix allows one to assess the effects of closure because the open variables are defined to have zero covariances (they are uncorrelated). Closure of this matrix yields covariances due soley to closure itself, so the correlations of the closed hypothetical matrix can serve as null values against which observed correlations (that is, Purdy's) can be tested.

Table 5.--Statistics describing set of 12 modal carbonate variables (recomputed to 100 percent)

Summary statistics

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains
Mean	.3174	4.749	2.962	2.369	.9848	5.488	16.56	5.107	13.24	1.245	30.63	16.35
Variance	.9923	48.78	27.53	7.940	13.57	42.47	438.5	28.70	252.6	5.507	1061.	205.7
Standard deviation	.9961	6.984	5.247	2.818	3.684	6.517	20.94	5.357	15.89	2.347	32.58	14.34
Coefficient of variation	3.1384	1.4705	1.7711	1.1895	3.7407	1.1874	1.2647	1.0489	1.2004	1.8846	1.0636	.87718
Maximum	9.158	47.72	35.79	15.94	32.87	43.37	79.34	31.46	59.07	14.39	99.10	64.19
Minimum	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
Range	9.158	47.72	35.79	15.94	32.87	43.37	79.34	31.46	59.07	14.39	99.10	64.19
Percent Variance	.04651	2.2862	1.2902	.37215	.63605	1.9904	20.551	1.3451	11.837	.25813	49.745	9.6420

Variance - covariance matrix

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains
Coralline algae	.99227	2.3653	.00076	.40470	2.3745	1.0673	-3.7658	-.050709	-.98443	.92649	-6.3063	2.9760
Halimeda		48.777	5.1807	5.7551	8.5013	18.971	3.1347	8.1967	-22.369	2.7415	-79.844	-1.4101
Peneroplidae			27.527	9.2069	.10897	15.558	33.083	8.8887	-21.289	.97021	-67.219	-12.016
Other forams				7.9400	.12902	8.4199	8.1578	4.8109	- 5.7921	1.5113	-46.537	5.9931
Corals					13.571	9.3060	-12.215	1.3964	- 5.4817	2.0831	-24.242	4.4692
Molluscs						42.467	15.607	8.1456	-14.845	2.2440	-108.54	1.5991
Fecal pellets							438.46	60.022	-136.90	-9.9426	-260.78	-134.87
Mud aggregates								28.698	-28.232	.44089	-73.016	-19.300
Grapestone									252.56	8.1491	-164.79	139.97
Organic aggregates										5.5075	-25.792	11.161
Oolites											1061.4	-204.29
Cryptocrystalline grains												205.72

Correlation coefficient matrix

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains
Coralline algae	1.0000	.33999	.00015	.14418	.64707	.16442	-.18054	-.00950	-.06219	.39632	-.19433	.20830
Halimeda		1.00000	.14138	.29244	.33043	.41682	.02143	.21908	-.20154	.16726	-.35092	-.01408
Peneroplidae			1.00000	.62277	.00563	.45503	.30114	.31625	-.25532	.07879	-.39326	-.15968
Other forams				1.00000	.01243	.45853	.13826	.31871	-.12934	.22854	-.50694	.14829
Corals					1.00000	.38765	-.15835	.07076	-.09363	.24095	-.20200	.08451
Molluscs						1.00000	.11437	.23333	-.14335	.14673	-.51124	.01711
Fecal pellets							1.00000	.53508	-.41138	-.20233	-.38227	-.44907
Mud aggregates								1.00000	-.33162	.03507	-.41838	-.25119
Grapestone									1.00000	.21850	-.31828	.61408
Organic aggregates										1.00000	-.33735	.33157
Oolites											1.00000	-.43720
Cryptocrystalline grains												1.00000

The simplest possible model of the hypothetical open matrix is that suggested by Chayes and Kruskall (1966). The means and variances of the variables of the open matrix are such that closure of the open matrix yields variables having means and variances equal to those in the observed (closed) matrix, that is, Purdy's matrix. The concern, of course, is to avoid attributing significance to correlations which reflect nothing more than numerical bias.

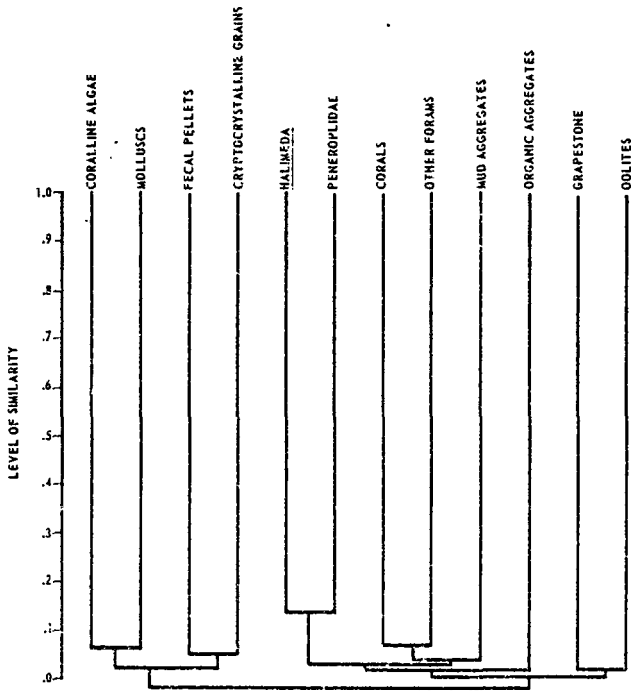
A program (Harbaugh and Bonham-Carter, 1970) designed to generate random numbers whose distributions are other than normal (the distributions are referred to as "empirical"; that is, they conform to the characteristics of the observed data rather than to a specific statistical distribution) was used to simulate a hypothetical open matrix of 12 uncorrelated variables measured on 1000 samples. The simulated variables are assigned means and variances such that closure of the open matrix yields variables having means and variances equal to those of Purdy's data. The results of the simulation are given in table 6; the approximation to Purdy's data is reasonably good. The r matrix for the simulated open matrix actually has mostly nonzero elements (the range in r values is 0.000 to 0.137) because of sample size and numerical consequences of the simulation.

For purposes of assessing the closure effect, the simulated open data were clustered using Pearson's r as the measure of similarity (fig. 6A) because this is the measure Purdy used. The dendrogram shows essentially no structure, which would be expected when the covariances of the variables equal zero. (The negligible structure which is apparent is a function of sample size.) Next, the simulated closed data were

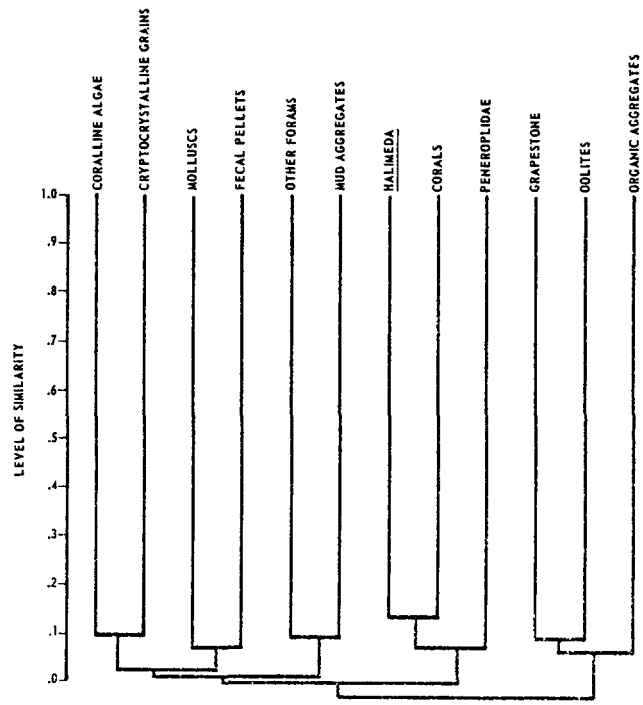
Table 6.--Comparison of means and variances of simulated and observed carbonate data sets

VARIABLE	OBSERVED MEAN	SIMULATED OPEN MEAN	SIMULATED CLOSED MEAN	OBSERVED VARIANCE	SIMULATED OPEN VARIANCE	SIMULATED CLOSED VARIANCE
Coralline algae	.32	.75	.87	.99	1.08	1.84
Halimeda	4.75	4.52	5.06	48.78	23.88	29.78
Peneroplidae	2.96	3.15	3.42	27.53	16.65	16.37
Other forams	2.37	2.60	2.95	7.94	6.17	9.26
Corals	.99	2.37	2.56	13.57	9.80	9.26
Molluscs	5.49	5.32	5.88	42.47	27.76	33.22
Fecal pellets	16.56	14.34	13.51	438.46	297.20	182.05
Mud aggregates	5.11	5.40	6.03	28.70	28.54	37.93
Grapestone	13.24	13.91	13.22	252.56	253.77	188.50
Organic aggregates	1.25	3.68	4.28	5.51	2.95	7.44
Oolites	30.63	30.60	26.08	1061.40	988.43	424.45
Cryptocrystalline grains	16.35	15.52	16.13	205.72	164.92	172.97

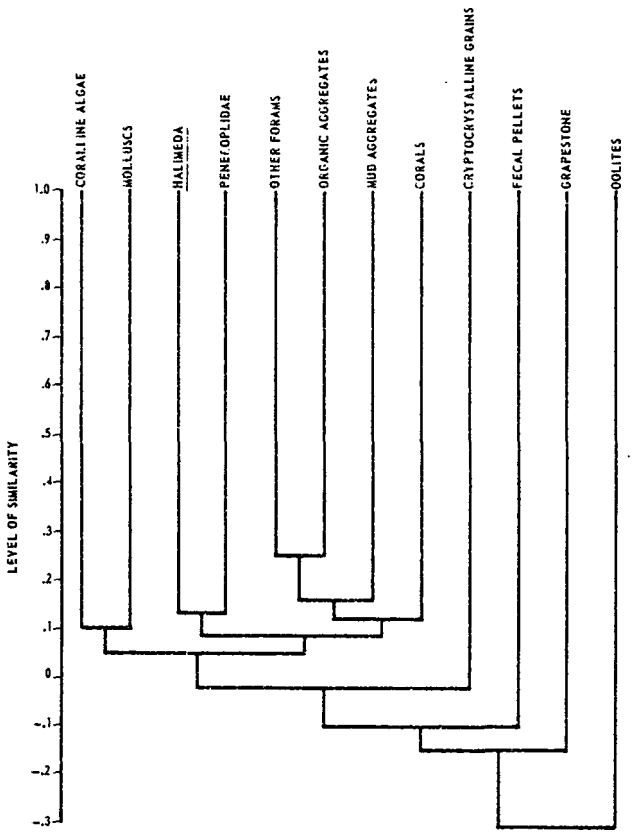
FIGURE 6. -- DENDROGRAMS OF SIMULATED CARBONATE DATA.



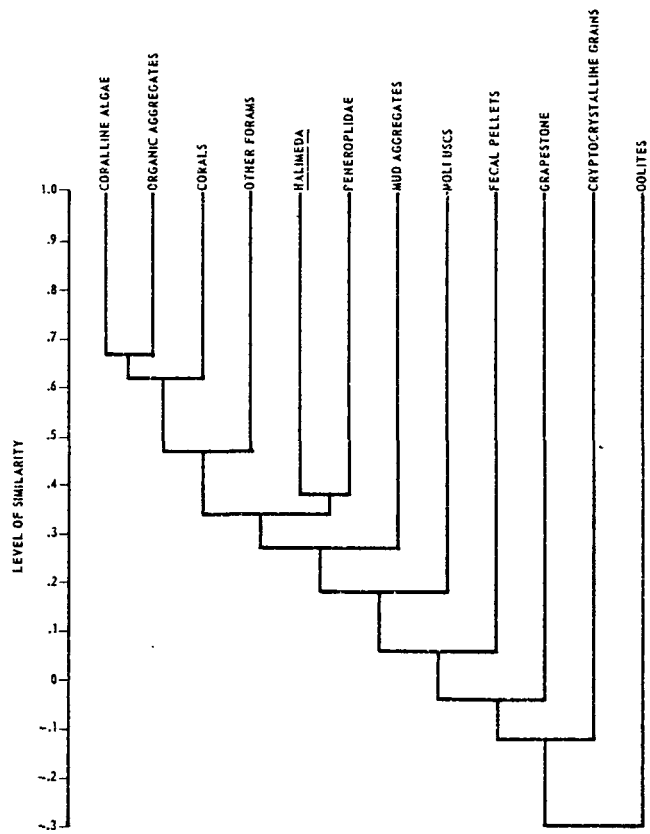
A. OPEN DATA; r AS MEASURE OF SIMILARITY



B OPEN DATA; RANK CORRELATION COEFFICIENT AS MEASURE OF SIMILARITY



C. CLOSED DATA; r AS MEASURE OF SIMILARITY



D. CLOSED DATA; RANK CORRELATION COEFFICIENT AS MEASURE OF SIMILARITY.

clustered, also using Pearson's r as the measure of similarity (fig. 6C). The two dendrograms (figs. 6A, C) are difficult to compare because the sequence in which variables enter the cluster are not the same, but closure has imposed at least some structure upon the dendrogram. Because closure of the simulated data has imposed so (relatively) little structure, one may assume that closure has not imposed much structure on the observed data (as the simulated data was designed to conform to the observed data). Therefore, for all practical purposed, one may conclude that the relationships apparent in Purdy's Λ cluster of the observed data are real and are not significantly affected by closure.

However, because the simulated data are of a nonuniform distribution, a nonparametric measure of similarity would be more appropriate. Therefore both Spearman's and Kendall's rank coefficients for the simulated open (table 7) and closed (table 8) data were computed. Figure 6 (B,D) shows the clusters formed using Spearman's coefficients. The cluster of the open data (fig. 6B) shows negligible structure, as expected. The cluster of the closed data (fig. 6D) suggests, at least within the ability to compare, that at least one of the observed reaction groups (coralline algae and corals) (fig. 5B) may not be statistically justifiable, regardless of what one might intuitively expect. The point, however, is that closure can induce correlation of the rank type; previous work has examined only the product-moment correlation.

Table 7.--Rank-correlation-coefficient matrix for simulated open carbonate data (Spearman's coefficient is in upper-right half of matrix; Kendall's coefficient is in lower-left half)

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains
Coralline algae	1.000	-.007	-.093	.040	-.020	.016	-.026	-.039	-.007	-.001	.013	.095
Halimeda	-.006	1.000	.118	.051	.135	.065	-.030	-.002	-.152	-.084	-.051	.132
Peneroplidae	-.087	.106	1.000	-.066	.013	.045	-.010	-.010	-.037	-.059	.058	.079
Other forams	.038	.046	-.061	1.000	.115	.077	.010	.090	-.057	-.046	-.040	.038
Corals	-.020	.125	.012	.108	1.000	-.097	-.068	.028	-.019	-.021	.102	.031
Molluscs	.015	.058	.039	.069	-.089	1.000	.073	-.067	.016	-.106	-.096	.081
Fecal pellets	-.024	-.027	-.010	.009	-.062	.063	1.000	-.002	.037	.006	.022	.019
Mud aggregates	-.035	-.002	-.009	.081	.026	-.058	-.001	1.000	-.010	.050	-.026	-.011
Grapestone	-.006	-.131	-.032	-.050	-.016	.013	.031	-.008	1.000	.053	.086	-.064
Organic aggregates	-.001	-.078	-.054	-.042	-.020	-.097	.005	.046	.047	1.000	.070	-.037
Oolites	.001	-.042	.048	-.033	.089	-.079	.018	-.020	.069	.062	1.000	.036
Cryptocrystalline grains	.082	.110	.069	.030	.027	.066	.015	-.008	-.052	-.033	.027	1.000

Table 8.--Rank-correlation-coefficient matrix for simulated closed carbonate data (Spearman's coefficient is in upper-right half of matrix; Kendall's coefficient is in lower-left half)

	Coralline algae	Halimeda	Peneroplidae	Other forams	Corals	Molluscs	Fecal pellets	Mud aggregates	Grapestone	Organic aggregates	Oolites	Cryptocrystalline grains
Coralline algae	1.000	.394	.339	.508	.631	.380	.118	.345	.118	.670	-.331	.120
Halimeda	.371	1.000	.379	.332	.441	.256	.011	.206	-.110	.365	-.312	.144
Peneroplidae	.344	.322	1.000	.230	.412	.228	-.005	.192	-.038	.355	-.241	.055
Other forams	.468	.287	.223	1.000	.468	.258	.076	.311	.015	.430	-.311	.060
Corals	.639	.414	.396	.445	1.000	.239	.033	.367	.064	.608	-.273	.054
Molluscs	.327	.218	.197	.222	.229	1.000	.091	.107	.023	.301	-.373	.051
Fecal pellets	.173	.050	.050	.109	.113	.099	1.000	.025	-.099	.145	-.383	-.128
Mud aggregates	.313	.181	.182	.258	.324	.102	.056	1.000	.043	.362	-.268	-.004
Grapestone	.176	-.017	.022	.061	.139	.047	-.019	.055	1.000	.158	-.246	-.202
Organic aggregates	.621	.321	.332	.385	.580	.256	.180	.312	.181	1.000	-.292	-.001
Oolites	-.160	-.187	-.131	-.173	-.132	-.236	-.216	-.155	-.148	-.137	1.000	-.308
Cryptocrystalline grains	.134	.123	.063	.072	.079	.050	-.055	.010	-.119	.042	-.193	1.000

Principal components analysis and reaction groups

As mentioned earleir, cluster analysis is not an analytically rigorous technique. Principal components analysis (PCA), although not a statistical procedure, is a mathematical manipulation by thich the "redundancy" in a set of variables becomes apparent. PCA was applied to the carbonate data with the intent that the technique might be a more quantitative discriminator--relative to cluster analysis--of the "real" groupings of the variables.

Any set of correlated variables can be transformed into a set of uncorrelated variables by a linear transformation that can be interpreted geometrically as a (rigid) rotation of the coordinate system to a position that concentrates as much as possible of the total variability of the data into a single new variable. The origin of the new coordinate system coincides with the means of the variables; the sum of the squared distances from the data points to the new ordinate is a minimum and the sum of the squared distances to the new abscissa is a maximum. The axis along which the variability is maximized is the major axis of an ellipsoid (hyperellipsoid in the case of multivariate data), which represents the new variable of maximum variance. The second new variable accounts for as much as possible of the remaining variability, is represented by the largest minor axis of the hyperellipsoid, and is therefore uncorrelated with the first new variable. Other uncorrelated

variance-maximizing variables may be formed until almost all the variability is accounted for by a few new variables. As a multivariate technique, the objective of PCA is to determine which of all the original variables are algebraically independent--that is, how many of the original variables actually represent the total amount of information; some variables may be simply linear combinations of another. (Afifi and Azen (1972) present a very thorough discussion of principal components analysis.)

Several PCA options were performed on three sets of data: Purdy's observed data, the simulated open data, and the simulated closed data. The results (table 9) show that PCA transformation has not been an efficient data reducer. Cases E, F, and H of table 9 show all the transformed variables have roughly equivalent values of percent variance (an exception is in case H wherein one variable has zero percent variance because the data of this analysis are closed). This implies that there are no new variables which can explain the total variability of the original data any more efficiently than the original data themselves explain it. In other cases where a few new variables have "substantially positive" values of percent variance, it is apparent that the distribution of the variance among the transformed variables is no more efficient than it was among the original untransformed variables. Because only a few of the original variables account for about 90 percent of the

Table 9.--Summary of principal components analysis (PCA) of carbonate modal data

Variable	Percent variance of untransformed variable		Percent variance of transformed variable		
			PCA option: variance/covariance matrix; original data	PCA option: r matrix; standardized data	PCA option: Spearman rank correlation coefficient; standardized data
Purdy's observed data (203 x 13)					
Coralline algae	.03%		54.2%	30.5%	38.7%
Halimeda	1.45		30.6	21.8	23.1
Peneroplidae	.79		6.5	13.6	11.3
Other forams	.24		4.2	7.6	5.8
Fecal pellets	15.26				
Oolites	40.97	80%			
Sample fraction < 1/8 mm	22.16		A	B	C
Simulated open data (1000 x 12)					
Coralline algae	.06%		54.4%	10.5%	12.0%
Halimeda	1.31		16.4	9.7	10.5
Peneroplidae	.91		13.9	9.0	9.8
Other forams	.34		9.0	8.8	9.3
Fecal pellets	16.32			for all remaining variables, percent variance is between 7.0 and 8.7	for all remaining variables, percent variance is between 6.0 and 8.9
Grapestone	13.93	84%			
Oolites	54.27		D	E	F
Simulated closed data (1000 x 12)					
Coralline algae	.17%		45.1%	17.7%	32.8%
Halimeda	2.68		19.8	11.3	10.9
Peneroplidae	1.47		18.0	10.0	10.5
Other forams	.83		7.5	9.5	8.6
Fecal pellets	16.36			for all remaining variables, percent variance is between 6.0 and 8.5, except for the last value of percent variance, which is 0.0	
Grapestone	16.94				
Oolites	38.13	87%			
Crytocrystalline grains	15.54		G	H	I

total variability:

	203x12	203x13
fecal pellets	20.6%	15.3%
grapestone	11.9	9.4
oolites	50.0	41.0
cryptocrystalline grains	9.6	7.1
weight percent < 1/8 mm		22.2
	92.1%	95.0%

perhaps the reaction groupings are actually a function of the large variances of these few variables. These few variables so dominate the total variability that perhaps groups fall out such that each group is defined by one of the dominant variables. If so, analytical methods couldn't help but reclaim these groups. In fact, such an interpretation would not be generally inconsistent with the reaction groups defined by Purdy. An exception would be the group coralline-algae-plus-corals, which has already been shown to be suspect because of closure.

Summary

Although r is a dubious choice for measure of similarity for Purdy's data, it yields a dendrogram comparable to that produced by rank correlation coefficients. At least in this case, cluster analysis grouped together similarly behaving variables regardless of the measure of similarity or the distribution of the data.

Examination of the effect of closure on Purdy's analysis reveals that at least one of the reaction groups may not be statistically justifiable. Moreover, after closure rank correlation coefficients are nonzero. Closure therefore induces correlation of the rank type as well as of the product-moment type.

PCA did not efficiently reduce the data and therefore may suggest that the reaction groups are a function of the large variances of a few variables.

STRATIGRAPHIC THICKNESS DATA IN RATIO FORM

Stratigraphic studies are replete with numerical applications, but the stratigraphic literature shows little inquiry into potential numerical pitfalls. This section describes how correlation bias resulting from ratio formation can affect stratigraphic investigations.

Stratigraphic thickness data is used in the form in which it is collected to prepare isopach and isolith maps, but these maps generally are tools for, not the objective of, stratigraphic investigations. Depending on the objective of the study, facies map design will probably require a transformation of absolute thickness measurements to percentages or ratios. Krumbein (1956, p. 2163) stated that "the selection of the method of facies expression (percentages, ratios, et cetera) [is among] the primary geologic considerations in lithofacies map design".

Stratigraphic maps may be of the contour type, which is suitable for studying rates of change in lithologic composition and for predictive purposes (Krumbein and Sloss, 1963). Contour-type maps could be prepared from percentage data, which would show the proportional thickness of lithologic components relative to the total thickness of the section, or from ratio data, which would contrast one lithology to another and suggest interrelationships among the components.

Krumbein (1962) observed that several of the maps prepared during a stratigraphic study resemble each other in contour pattern. This suggested to him that the geometric and compositional attributes of the unit being mapped were somehow "interlocked", that is, that an individual map was in part repeating the information shown in another. In citing the work

of Chayes (1960, 1962), Krumbein noted that "data interlock" is at least partly dependent on the manner in which the numerical data are expressed, and that the "built in" numerical relationships among variables have direct bearing on the contour patterns of facies maps. Krumbein also suggested that the correlation coefficient is an indicator of the degree of linear data interlock and contended that it may be used to facilitate the most efficient selection of maps to be prepared in a stratigraphic study (Krumbein, 1962, p. 2233). Krumbein seems to be suggesting that the numerical bias in percentage and ratio correlation coefficients can be turned to the advantage of the stratigrapher. The analysis in this section should indicate that it probably cannot.

To the extent that the correlation coefficient can suggest possible relations among geologic variables, its use is certainly appropriate. "This is an empirical procedure, fully justifiable in the early stages of geologic analysis, and . . . may be quite effective in 'sorting out' the interrelationships among . . . variables" (Krumbein and Graybill, 1965, p. 236). For example, in some stratigraphic models, such as the clastic wedge, "the sand thickness commonly shows a moderate degree of positive correlation with total unit thickness. In other basin models, in which carbonates or evaporites are dominant, the sand thins as the unit thickens toward the basin center, giving rise to negative correlation" (Krumbein and Graybill, 1965, p. 236). Depositional models described by Krumbein and Sloss (1963) corroborate these observations and suggest others (for example, a positive correlation between shale thickness and total unit

thickness at the cratonic border of marginal basins or at the hinge lines of interior basins).

The extent to which these and similar observations may be biased by the data treatment which produces them will be analyzed using an example from the literature. The data is Imbrie's (1963) and has been cited by Krumbein (1962) and by Krumbein and Graybill (1965) for various analytical purposes. The data (table 10, taken from Krumbein and Graybill, 1965) are from 31 wells in western Kansas and southeastern Colorado, drilled in Upper Permian rocks including sandstone, shale, evaporites, and carbonates (the variable "clastics" as used in this thesis is the sum of sand and shale; "nonclastics" is the sum of carbonate and evaporite). The wells were chosen in an area where the lithologic components thicken toward a center, as is characteristic of sedimentary basins (Krumbein, 1962, p. 2235). Imbrie's data are used in this section to examine correlation bias resulting from ratio formation.

Some ratios which commonly appear in stratigraphic studies are:

sand/shale	
clastics/nonclastics	(the classic ratio)
evaporites/carbonates	(the evaporite ratio)
nonclastics/sand	
nonclastics/shale	

The literature (e.g., Krumbein, 1962; Krumbein and Graybill, 1965; Krumbein and Sloss, 1963) shows that these ratios and the lithologic variables which form them have been correlated in stratigraphic work as

Table 10.--Stratigraphic thickness data (from Krumbein and Graybill, 1965, p. 372)

Total thickness	Sand	Shale	Nonclastics	Carbonates*	Evaporites
845	266	350	229	24	205
906	337	432	137	60	77
844	451	511	82	42	40
447	293	116	38	12	26
1,001	548	450	205	17	186
955	273	436	223	41	182
374	240	110	24	24	0
508	365	148	95	20	75
640	224	304	112	14	98
614	255	372	87	28	59
915	295	555	265	43	222
1,159	179	643	317	20	297
792	237	341	124	39	85
464	104	242	118	18	100
1,118	180	568	370	0	370
1,224	207	758	259	11	248
1,204	277	610	317	10	307
1,144	320	520	314	12	302
1,048	562	310	176	12	164
1,162	130	659	373	13	360
1,005	224	542	237	21	216
721	229	400	92	12	80
775	223	477	75	28	47
1,025	295	501	227	18	209
1,114	246	528	340	32	308
955	267	502	186	24	162
532	157	238	157	0	157
562	120	316	126	0	126
1,005	271	637	97	8	89
530	30	461	39	0	39
1,126	270	558	298	68	230

* Entries equal to 0 were changed to 1 to avoid division by 0 when forming the evaporite ratio.

shown in table 11. The pairs indicated by 0 in table 11 fall into the category of "part-whole correlation" (Chayes, 1971, p. 25-26).

Part-whole correlation can be defined as follows. Suppose X is a vector of random variables $[x_1, x_2]$ and that Q_i is defined:

$$Q_i = x_1 + x_2 \quad \text{(the whole)};$$

Q_j is defined:

$$Q_j = x_1 \qquad \text{(the part)}$$

The part-whole correlation, r_{ij}, to be expected between Q_i and Q_j when x_1 and x_2 are random variables is equal to the ratio of the standard deviation (s) of the part to the standard deviation of the whole:

$$r_{ij} = \frac{s_{x_1}}{s_{x_1 + x_2}}$$

This equation and equations 2-6 in table 2 show that part-whole correlation and ratio correlation applied to stratigraphic data will be biased; the null value of r will not be zero. Analysis of Imbrie's data with respect to some of the correlated pairs of variables specified in table 11 will attempt to assess the magnitude of the bias and account for it in subsequent data interpretation.

Simulation (Harbaugh and Bonham-Carter, 1970) is used to establish the null correlations for the stratigraphic data. The data of Imbrie (as given by Krumbein and Graybill, 1965) and some derivative variables (computed for this thesis) are described by the summary statistics and the correlation coefficient matrix in table 12. (According to a chi-square goodness-of-fit test, all of Imbrie's variables fit a normal

Table 11.--Stratigraphic ratio correlations (X) and part-whole correlations (0)

	Total thickness	Sand	Shale	Clastics	Carbonates	Evaporites	Nonclastics	Sand/shale	Clastic ratio	Evaporite ratio	Nonclastics/sand	Nonclastics/shale
Total thickness								X	X	X		
Sand	0							X				
Shale	0							X				
Clastics	0	0	0						X			
Carbonates												
Evaporites	0											
Nonclastics	0				0	0		X	X		X	X
Sand/shale									X	X		X
Clastic ratio										X		
Evaporite ratio												
Nonclastics/sand												
Nonclastics/shale												

Table 12.--Correlation-coefficient matrix and summary statistics for observed stratigraphic thickness data

	Total Thickness	Sand	Shale	Nonclastics	Carbonates	Evaporites	Clastics	Sand/shale	Clastic ratio	Evaporite ratio	Nonclastics/sand	Nonclastics/shale
Total Thickness	1.00											
Sand	.243	1.00										
Shale	.887	-.123	1.00									
Nonclastics	.844	-.035	.690	1.00								
Carbonates	.141	.454	-.055	.058	1.00							
Evaporites	.818	-.108	.697	.987	-.101	1.00						
Clastics	.948	.372	.875	.629	.170	.599	1.00					
Sand/shale	-.517	.506	-.748	-.500	.124	-.519	-.453	1.00				
Clastic ratio	-.570	.036	-.403	-.787	-.066	-.774	-.359	.488	1.00			
Evaporite ratio	.082	-.335	.113	.297	-.432	.366	-.058	-.228	-.242	1.00		
Nonclastics/sand	.508	-.556	.595	.767	-.233	.802	.281	-.603	-.552	.437	1.00	
Nonclastics/shale	.353	.087	-.048	.714	-.099	.696	.088	-.035	-.796	.276	.462	1.00
Mean	860.6	247.3	428.8	184.4	21.8	162.8	676.2	.745	5.08	29.74	.859	.427
Standard deviation	254.2	85.6	164.1	104.1	16.4	104.5	175.5	.607	3.29	70.67	.602	.167
Coefficient of variation	.295	.346	.383	.565	.753	.642	.260	.814	.648	2.38	.700	.392
Variance	64610.0	7322.0	26940.0	10840.0	269.0	10910.0	30790.0	.368	10.8	4994.0	.362	.028
Percent variance	42.2	4.7	17.2	6.9	.2	7.0	19.7	.0002	.007	3.19	.0002	.00002

distribution.) These summary statistics (table12) are used to assign values to the parameters (means and variances) of a simulated data set. The simulated matrix has 1000 samples; the variables are the same as those in table 12 and have the same means and variances, but the covariances of the simulated data are defined to be zero. Because the covariances of the simulated data are zero, any correlations apparent in the simulated data (table 13) are due solely to the data treatment and are not inherent in the variables. The simulated correlation coefficients, therefore, serve as the null values for the data in ratio or part-whole form.

Looking only at observed values of r which are significant and referring to the variable pairs specified in table 11, one may make some interesting observations (table 14). In all but a few cases, what appear to be highly significant correlations actually approximate the corresponding null correlations. The implication is, of course, that, at least in this case, these ratio correlations cannot be useful in inferring geologic relationships among the variables. One's use of them, for example, as criteria for selecting maps to be incorporated into a stratigraphic study is ill-conceived on the basis of these results.

On the other hand, what appears to be a relatively weak association (-.359), which might likely be ignored, is actually significantly departing from randomness (null value = .315).

Table 13.--Null correlation coefficients, derived by simulation, for stratigraphic thickness data

	Sand	Shale	Nonclastics	Carbonates	Evaporites	Clastics	Total thickness	Sand/shale	Clastic ratio	Evaporite ratio	Nonclastics/sand	Nonclastics/shale
Sand	1.00											
Shale	.035	1.00										
Nonclastics	-.043	.009	1.00									
Carbonates	-.013	.052	.012	1.00								
Evaporites	-.010	.025	.025	.042	1.00							
Clastics	.499	.884	-.011	.039	.017	1.00						
Total thickness	.357	.693	.439	.121	.476	.768	1.00					
Sand/shale	.502	-.695	-.003	-.034	-.021	-.368	-.294	1.00				
Clastic ratio	.175	.269	-.751	.025	-.046	.315	-.107	-.112	1.00			
Evaporite ratio	-.005	-.022	-.007	-.495	.431		.141	.019	-.005	1.00		
Nonclastics/sand	-.564	-.039	.464	-.012	-.025	-.298	-.037	-.271	-.398	.005	1.00	
Nonclastics/shale	-.039	-.582	.651	-.026	-.023	-.523	-.127	.539	-.596	.010	.318	1.00
Mean	247.8	443.3	184.9	23.0	161.5	691.0	1060.4	.67	5.6	14.1	.97	.50
Standard deviation	82.4	152.6	100.6	15.6	103.4	175.9	230.5	.44	4.5	21.0	1.2	.42
Coefficient of variation	.33	.34	.54	.68	.64	.25	.22	.67	.81	1.5	1.25	.84
Variance	6789.8	23286.8	10120.4	243.1	10700.5	30940.8	53130.3	.19	20.3	441.0	1.44	.18
Percent variance	5.0	17.2	7.5	.18	7.9	22.8	39.2	0.0	.02	.003	0.0	0.0

Table 14.--Matrix of $\frac{r_{observed}}{r_{null}}$ for stratigraphic thickness data

	Total Thickness	Sand	Shale	Clastics	Carbonates	Evaporites	Nonclastics	Sand/shale	Clastic ratio	Evaporite ratio	Nonclastics/sand	Nonclastics/shale
Total thickness		.243 .357	.887 .693	.948 .768		.818 .476	.844 .439					
Sand				.372 .499				.506 .502				
Shale				.875 .884				-.748 -.695				
Clastics									-.359 .315			
Carbonates												
Evaporites												
Nonclastics												
Sand/shale									-.787 -.753		.767 .464	.714 .651
Clastic ratio												
Evaporite ratio												
Nonclastics/sand												
Nonclastics/shale												

Simulated vs. approximated null values

In the foregoing analysis simulation was used to derive the null values for the ratio and part-whole correlations because the coefficients of variation (C) of the variables were too large to permit the use of Chayes' (1971) approximations (table 2). According to Chayes, for C larger than 0.15 the differences between the approximated and the simulated null

correlations may be large, but for $C \leq 0.15$ the differences are negligible and the approximations are adequate. "Before the development of numerical simulation, this inadequacy [of the approximations] was of course critical . . . but it is now a simple matter to run a simulation experiment on variables characterized by any set of means, variances and covariances" (equal to zero for simulation of null correlations) (Chayes, 1971, p. 15, 17).

All the stratigraphic variables in table 12 have C larger than 0.15. The correspondence between the simulated null values and those obtained by Chayes' approximations is of some academic, if not practical, interest.

Table 15 shows the same correlated variables that were examined in the foregoing analysis, but the values in the table represent two null correlations, that obtained by simulation and that obtained by Chayes' approximation formulas. As expected, where the C's of the variables being correlated are both "relatively small", the correspondence between the nulls is good (differing only in the second or third decimal place). Good correspondence, however, occurs even when one C is very large. No meaningful relationship could be discerned between the magnitude of the difference between the nulls and the magnitude of the C's of the variables being correlated. It is apparent, though, that the part-whole correlations show consistently good agreement regardless of the C's, whereas the ratio correlations show consistently bad agreement. This reflects the fact that the part-whole correlation is simply the ratio of the standard deviations of the variables, and the standard deviations of the observed and simulated variables are approximately equal.

Table 15.--Matrix of $\frac{r_{\text{simulated null}}}{r_{\text{approximated null}}}$ (upper-right half of matrix) and the difference between the two nulls (lower-left half of matrix; difference is positive (+) if $r_{sim.} > r_{approx.}$ and difference is negative (-) if $r_{sim.} < r_{approx.}$)

	Total thickness	Sand	Shale	Clastics	Carbonates	Evaporites	Nonclastics	Sand/shale	Clastic ratio	Evaporite ratio	Nonclastics/sand	Nonclastics/shale
Total thickness		.357 / .357	.693 / .662	.768 / .763		.476 / .449	.439 / .437					
Sand	.000			.499 / .469				.502 / .697				
Shale	.031							-.695 / -.718				
Clastics	.005	.030	.017						.315 / .420			
Carbonates												
Evaporites	.027								-.753 / -.907		.464 / .853	.651 / .828
Nonclastics	.022											
Sand/shale		-.195	-.023									
Clastic ratio				-.105			-.154					
Evaporite ratio												
Nonclastics/sand												
Nonclastics/shale												
Coefficient of variation (observed)	.295	.346	.383	.260	.753	.642	.565	.814	.648	2.38	.700	.392

Summary

Because the r matrix reflects the numerical bias induced by ratio formation (data treatment which is common in stratigraphic studies), what appear to be highly significant correlations actually approximate the corresponding null correlations. Consequently, it is conceivable that depositional models may be misinterpreted and that use of the observed correlations as criteria for selecting maps to be incorporated into a stratigraphic investigation is not statistically justifiable.

Correspondence between simulated null values and those obtained by Chayes' (1971) approximations is erratic, and the relationship between this correspondence and the C's of the variables being correlated could not be determined—and is probably not of practical import due to readily available simulation experiments.

GEOCHEMICAL DATA

Standardized vs. nonstandardized principal-components-analysis scores

PCA scores may be correlated or uncorrelated depending on whether standardized or nonstandardized data is used to compute them. How, numerically, the matrices of scores differ should provide insight into why, given the same eigenvalues, scores produced from one kind of data are correlated and those produced from another kind are not. This problem is examined using a set of geochemical analyses.

The summary statistics for chemical analyses (weight percents) of the oxides of 28 volcanic rocks from Gough Island are given in table 16. These 28 analyses were included in a larger data set analyzed (LeMaitre, 1968) by PCA for the purpose of determining differentiation trends. PCA is a mathematical manipulation whereby data is transformed to a new coordinate system; by definition the transformation produces uncorrelated variables. The objective of PCA is "to define a new reference system in which the total variance of the original data is preserved but the covariance is eliminated" (Trochimezyak and Chayes, in press). One should not, therefore, expect "trends" to be apparent in plots of scores derived from PCA.

This suggests that the use of PCA may be counterproductive — that is, will produce correlated new variables (scores) — under some circumstances. Trochimczyak and Chayes (in press) have shown under which circumstances (data treatment) the transformed variables are correlated.

Table 16.--Summary statistics describing chemical analyses of Gough Island volcanic rocks

	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
Mean	54.2	2.0	17.4	3.0	5.2	.1	4.1	5.4	4.4	4.0	.3
Variance	29.1	1.3	4.7	2.4	5.5	.003	18.3	9.1	2.2	3.3	.02
Standard deviation	5.4	1.2	2.2	1.6	2.3	.056	4.3	3.0	1.5	1.8	.14
Coefficient of variation	.1	.6	.1	.5	.5	.4	1.1	.6	.3	.5	.6
Percent variance	38.3	1.8	6.2	3.2	7.2	.004	24.1	12.0	2.8	4.4	.02
Range	15.8	5.3	11.3	6.7	8.2	.3	19.7	8.4	6.0	5.5	.5

Of the several options which extract principal components and then transform the original data into scores, some which are compatible with the objective of PCA are:

1) the use of the r matrix to extract principal components and the use of standardized data (Z-scores of original data) to compute scores;

2) the use of the variance-covariance matrix to extract principal components and the use of the original data to compute scores.

Other options, such as

3) a combination of the r matrix and the original data, or

4) a combination of the variance-covariance matrix and Z-scores

are unsatisfactory methods of PCA because they result in correlated scores.

All four of these PCA options were computed for the Gough Island data in order to examine the results with respect to those which PCA, by definition, should produce. The PC analyses are available in the geology department of the University of Houston. The Gough Island results from options 1 and 3 (and options 2 and 4) differ only by the new variables; therefore plots of the scores differ. All other analytical results are the same for each pair of options.

The plots of the scores are consistent with the results about correlation that Trochimczyak and Chayes (in press) predict. Furthermore, whenever standardized data is used to compute scores, the plot of the scores centers around zero for both variables. This is to be expected, as Z-scores measure the "distance" of x from $\bar{x}$, so the mean of the Z-scores is zero.

The numerical difference (correlated vs. uncorrelated variables) which the various PCA options have produced cannot be explained simply by inspection of the plots. It is apparent, however, that when standardized data are used to compute scores (regardless of the matrix used to extract principal components), the scatter of points is noticeably compressed. Another observation is that, again regardless of the matrix used, absolutely no overlap of points occurs between plots of scores from standardized vs. original data.

Although inspection failed to yield any substantive insight into why, numerically, the plots differ, a more systematic analysis might succeed. As the matrices of the Gough Island PC analyses are too unwieldy to manipulate, a 5 x 3 data matrix was "simulated" such that the first variable has a mean many times larger than the second variable and not as many times larger than the third variable. The eigenvectors (in the working model, a 3 x 3 subset from the Gough Island r-matrix PC analyses) used to compute the scores are the same for both original and standardized data.

The appendix contains the matrix manipulations which were intended to reveal how, numerically, the matrices of PCA scores, derived from standardized vs. original data, differed; this, then, might provide a clue as to why, given the same eigenvectors, scores produced from one kind of data were correlated and those produced from another kind were not.

To define a term used in the following discussion, the product of two matrices is a matrix whose element in the i^{th} row and j^{th} column is the sum of what will be referred to as the "cross products" of elements in the i^{th} row and j^{th} column of the factors.

Inspection of any data matrix of chemical analyses and the corresponding matrix of Z-scores reveals an obvious difference between the two: about 50 percent of the elements in the matrix of Z-scores are negative. The negative elements represent the original values of the variable which are less than the mean of the variable. This difference carries through to the matrix of cross products (appendix, p. 1-3, circled elements). Where the value of the original data is less than the mean of the variable, the sign of the cross product will change when the data is standardized. No sign change occurs when the value of the original data is greater than $\bar{x}$. This suggests that the values of $\bar{x}$ are affecting the numerical aspect of the matrix of PCA scores, and page 4 of the appendix shows how. If A is the sum of the cross products (that is, if A is the matrix of PCA scores) when original data are used (appendix, p. 2) and if B is the sum of the cross products when Z-scores are used (appendix, p. 3), and if C is the row vector whose components are the sums of cross products between the row vector of means of the original variables $[\bar{x}_1, \bar{x}_2, \bar{x}_3]$ and the matrix of eigenvectors (appendix, p. 4), then the sum of the i^{th} column of B and the i^{th} component of C is the i^{th} column of A.

This observation can be refined somewhat (appendix, p. 5) to see that for every element in the j^{th} column of the matrix of PCA scores, the q cross products (where q = the number of variables) which are summed to

compute the element differ (with respect to Z-score factor vs. original-data factor) by the amount:

(eigenvector matrix component$_{1j}$) x ($\bar{x}_1$) for the first cross product

(component$_{2j}$) x ($\bar{x}_2$) second

(component$_{3j}$) x ($\bar{x}_3$) third

(component$_{qj}$) x ($\bar{x}_q$) q^{th}

The value of $\bar{x}$'s, then, rather than of $(x-\bar{x})$ seems to be making the numerical difference in the matrices of PCA scores. How this information may be brought to bear on the correlation of scores is a subject for further investigation.

Summary

Under some circumstances (data treatment) PCA will produce correlated variables which, by definition, is indicative of analytical error. Matrix manipulation lent some insight into how, numerically, the matrices of PCA scores, derived from standardized vs. original data, differed. Apparently the value of the $\bar{x}$'s, rather than of $(x-\bar{x})$, seems to be making the numerical difference in the matrices of PCA scores.

CONCLUSIONS

This investigation has determined that some ramifications of data treatment in combination with quantitative analytical methods are:

--Closure can induce correlation of the rank type as well as of the product-moment type

--Dendrograms produced by the use of r can be generally reproduced by the use of rank correlation coefficients; at least in the case studied, cluster analysis will group together similarly behaving variables regardless of the measure of similarity or the distribution of the data

--PCA may not be any more efficient a reducer of data than a visual inspection of variances and part-whole correlations,

--Ratio correlations such as those used in stratigraphic studies, may be substantially biased and should not be used to infer geologic relationships among the variables; their use, for example, as criteria for selecting maps to be incorporated into a stratigraphic study may be ill-conceived

--The numerical difference between matrices of PCA scores derived from standardized vs. original data should provide a clue as to why, given the same eigenvectors, scores produced from one kind of data were correlated and those produced from another were not; apparently the value of the $\bar{x}$'s, rather than that of $(x-\bar{x})$, seems to be making the numerical difference.

REFERENCES CITED

Afifi, A.A., and Azen, S.P., 1972, Statistical analysis, a computer oriented approach: New York, Academic Press, 366 p.

Chayes, F., 1949, On ratio correlation in petrography: Jour. Geology, v. 57, p. 239-254.

_____, 1960, On correlation between variables of constant sum: Jour. Geophysical Research, v. 65, p. 4185-4193.

_____, 1962, Numerical correlation and petrographic variation: Jour. Geology, v. 70, p. 440-452.

_____, 1971, Ratio correlation: Chicago, Chicago Univ. Press, 99 p.

_____ and Kruskall, W., 1966, An approximate statistical test for correlations between proportions: Jour. Geology, v. 74, pt. 2, p. 692-702.

Cramer, H., 1955, The elements of probability theory: New York, John Wiley and Sons.

Davis, J.C., 1973, Statistics and data analysis in geology: New York John Wiley and Sons, 550 p.

Demirmen, F., 1976, RANK - A FORTRAN IV program for computation of rank correlations: Computers and geosciences, v. 1, p. 221-229.

Harbaugh, J.W., and Honham-Carter, G., 1970, Computer simulation in geology: New York, John Wiley and Sons, 575 p.

Imbrie, J., 1963, Factor and vector analysis programs for analyzing geologic data: Office Naval Res., Geography Branch, Tech. Rept. 6, ONR Task No. 389-135.

_____ and Purdy, E.G., 1962, Classification of modern Bahamian carbonate sediments, p. 253-272 in Classification of carbonate rocks - a symposium: Am. Assoc. Petroleum Geologists Mem. 1, 279 p.

Koch, G.S., Jr., and Link, R.F., 1971, Statistical analysis of geologic data, v. 2: New York, John Wiley and Sons, 438 p.

Kolstoe, R.H., 1973, Introduction to statistics for the behavioral sciences: Homewood, Illinois, The Dorsey Press, 383 p.

Krumbein, W.C., 1956, Regional and local components in facies maps: Am. Assoc. Petroleum Geologists Bull., v. 40, p. 2163-2194.

_____, 1962, Open and closed number systems in stratigraphic mapping: Am. Assoc. Petroleum Geologists Bull., v. 46, p. 2229-2245.

_____ and Sloss, L.L., 1963, Stratigraphy and sedimentation (2nd ed.): San Francisco: W.H. Freeman and Co., 660 p.

_____ and Graybill, F.A., 1965, An introduction to statistical models in geology: New York, McGraw-Hill Book Co., 475 p.

LeMaitre, R.W., 1962, Petrology of volcanic rocks, Gough Island, South Atlantic: Geol. Soc. America Bull., v. 73, p. 1339-1340.

_____, 1968, Chemical variation within and between volcanic rock series - a statistical approach: Jour. Petrology, v. 9, p. 220-252.

McCammon, R.B., 1969, Aspects of classification, in Models of geologic processes - an introduction to mathematical geology, AGI/CEGS short course, 7-9 November 1969, Philadelphia: Washington, D.C., Am. Geologic Institute, p. RM-C-1 to RM-F-6.

Neter, J. and Wasserman, W., 1974, Applied linear statistical models: Homewood, Illinois, Richard D. Irwin, Inc., 842 p.

Parks, J.M., 1966, Cluster analysis applied to multivariate geologic problems: Jour. Geology, v. 74, pt. 2, p. 703-715.

Pearson, K., 1896, On a form of spurious correlation which may arise when indices are used in the measurement of organs: Proc. Roy. Soc. (London), v. 60, p. 489-502.

Purdy, E.G., 1960, Recent calcium carbonate facies of the Great Bahama Bank: Ph.D. dissertation, Columbia Univ., 174 p.

Snedecor, G., 1956, Statistical methods: Ames, Iowa, Iowa State College Press.

Till, R., 1974, Statistical methods for the earth scientist: New York, John Wiley and Sons, 154 p.

Trochimczyak, J., and Chayes, F., in press, Some properties of principal component scores: Math. geology.

APPENDIX

Examination of matrices of PCA scores derived from standardized and nonstandardized data

ORIGINAL DATA (5 x 3)

$$\begin{bmatrix} 20 & 1 & 5 \\ 22 & 2 & 8 \\ 24 & 5 & 6 \\ 26 & 3 & 4 \\ 28 & 4 & 7 \end{bmatrix} \times \underset{(3 \times 3)}{\text{EIGENVECTORS}} \begin{bmatrix} E_{11} & -E_{12} & -E_{13} \\ -E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & -E_{33} \end{bmatrix} = \underset{(5 \times 3)}{\text{CROSS PRODUCTS}}$$

$$\begin{bmatrix} E_{11}(20)-E_{21}(1)+E_{31}(5) & -E_{12}(20)+E_{22}(1)+E_{32}(5) & -E_{13}(20)+E_{23}(1)-E_{33}(5) \\ E_{11}(22)-E_{21}(2)+E_{31}(8) & -E_{12}(22)+E_{22}(2)+E_{32}(8) & -E_{13}(22)+E_{23}(2)-E_{33}(8) \\ E_{11}(24)-E_{21}(5)+E_{31}(6) & -E_{12}(24)+E_{22}(5)+E_{32}(6) & -E_{13}(24)+E_{23}(5)-E_{33}(6) \\ E_{11}(26)-E_{21}(3)+E_{31}(4) & -E_{12}(26)+E_{22}(3)+E_{32}(4) & -E_{13}(26)+E_{23}(3)-E_{33}(4) \\ E_{11}(28)-E_{21}(4)+E_{31}(7) & -E_{12}(28)+E_{22}(4)+E_{32}(7) & -E_{13}(28)+E_{23}(4)-E_{33}(7) \end{bmatrix}$$

$$\bar{X} = \begin{bmatrix} 24 & 3 & 6 \end{bmatrix}$$

$$= \left[\begin{array}{ccc|ccc|ccc} + & - & + & - & + & + & - & + & - \\ + & - & + & - & + & + & - & + & - \\ + & - & + & - & + & + & - & + & - \\ + & - & + & - & + & + & - & + & - \\ + & - & + & - & + & + & - & + & - \end{array}\right]$$

STANDARDIZED DATA (Z-scores of original data)

$$\frac{1}{S}\begin{bmatrix} 20-24 & 1-3 & 5-6 \\ 22-24 & 2-3 & 8-6 \\ 24-24 & 5-3 & 6-6 \\ 26-24 & 3-3 & 4-6 \\ 28-24 & 4-3 & 7-6 \end{bmatrix}$$

$$\begin{bmatrix} - & - & - \\ - & - & + \\ 0 & + & 0 \\ + & 0 & - \\ + & + & + \end{bmatrix} \times \begin{bmatrix} \text{EIGENVECTORS} \\ \text{as} \\ \text{above} \end{bmatrix} = \frac{1}{S}\begin{bmatrix} E_{11}(-)-E_{21}(-)+E_{31}(-) & -E_{12}(-)+E_{22}(-)+E_{32}(-) & -E_{13}(-)+E_{23}(-)-E_{33}(-) \\ E_{11}(-)-E_{21}(-)+E_{31}(+) & -E_{12}(-)+E_{22}(-)+E_{32}(+) & -E_{13}(-)+E_{23}(-)-E_{33}(+) \\ E_{11}(0)-E_{21}(+)+E_{31}(0) & -E_{12}(0)+E_{22}(+)+E_{32}(0) & -E_{13}(0)+E_{23}(+)-E_{33}(0) \\ E_{11}(+)-E_{21}(0)+E_{31}(-) & -E_{12}(+)+E_{22}(0)+E_{32}(-) & -E_{13}(+)+E_{23}(0)-E_{33}(-) \\ E_{11}(+)-E_{21}(+)+E_{31}(+) & -E_{12}(+)+E_{22}(+)+E_{32}(+) & -E_{13}(+)+E_{23}(+)-E_{33}(+) \end{bmatrix}$$

$$= \frac{1}{S}\left[\begin{array}{ccc|ccc|ccc} - & + & - & + & - & - & + & - & + \\ - & + & + & + & - & + & + & - & - \\ 0 & - & 0 & 0 & + & 0 & 0 & + & 0 \\ + & 0 & - & - & 0 & - & - & 0 & + \\ + & - & + & - & + & + & - & + & - \end{array}\right]$$

ORIGINAL DATA (5 x 3)				EIGENVECTORS (3 x 3)				CROSS PRODUCTS (5 x 3)
20	1	5	×	.37	-.10	-.12	=	.37(20)-.34(1)+.28(5) -.10(20)+.22(1)+.39(5) -.12(20)+.08(1)-.06(5)
22	2	8		-.34	.22	.08		.37(22)-.34(2)+.28(8) -.10(22)+.22(2)+.39(8) -.12(22)+.08(2)-.06(8)
24	5	6		.28	.39	-.06		.37(24)-.34(5)+.28(6) -.10(24)+.22(5)+.39(6) -.12(24)+.08(5)-.06(6)
26	3	4						.37(26)-.34(3)+.28(4) -.10(26)+.22(3)+.39(4) -.12(26)+.08(3)-.06(4)
28	4	7						.37(28)-.34(4)+.28(7) -.10(28)+.22(4)+.39(7) -.12(28)+.08(4)-.06(7)

$$E = \begin{bmatrix} 7.4 - .34 + 1.4 & 2. + .22 + 1.95 & 2.4 + .08 - .3 \\ 8.1 - .68 + 2.2 & 2.2 + .44 + 3.12 & 2.6 + .16 - .5 \\ 8.9 - 1.7 + 1.7 & 2.4 + 1.1 + 2.34 & 2.9 + .4 - .4 \\ 9.6 - 1.02 + 1.1 & 2.6 + .66 + 1.56 & 3.1 + .24 - .24 \\ 10.4 - 1.4 + 1.96 & 2.8 + .83 + 2.73 & 3.4 + .32 - .42 \end{bmatrix}$$

SUM OF THE CROSS PRODUCTS = $A = \begin{bmatrix} 8.5 & .2 & -2.6 \\ 9.6 & 1.4 & -2.9 \\ 8.9 & 1.04 & -2.9 \\ 9.7 & -.4 & -3.1 \\ 11.0 & .8 & -3.5 \end{bmatrix}$ = PCA SCORES

A_1 A_2 A_3

	STANDARDIZED DATA (5 x 3)				EIGENVECTORS (3 x 3)				CROSS PRODUCTS (5 x 3)		
$\frac{1}{S}$	-4	-2	-1	×	.37	-.10	-.12	= $\frac{1}{S}$	.37(-4)-.34(-2)+.28(-1)	-.10(-4)+.22(-2)+.39(-1)	-.12(-4)+.08(-2)-.06(-1)
	-2	-1	2		-.34	.22	.08		.37(-2)-.34(-1)+.28(2)	-.10(-2)+.22(-1)+.39(2)	-.12(-2)+.08(-1)-.06(2)
	0	2	0		.28	.39	-.06		.37(0) -.34(2) +.28(0)	-.10(0) +.22(2) +.39(0)	-.12(0) +.08(2) -.06(0)
	2	0	-2						.37(2) -.34(0) +.28(-2)	-.10(2) +.22(0) +.39(-2)	-.12(2) +.08(0) -.06(-2)
	4	1	1						.37(4) -.34(1) +.28(1)	-.10(4) +.22(1) +.39(1)	-.12(4) +.08(1) -.06(1)

$$D = \left[\begin{array}{ccc|ccc|ccc} -1.7 & +.7 & -.3 & .4 & -.4 & -.4 & .5 & -.2 & +.06 \\ -.7 & +.3 & +.6 & .2 & -.2 & +.8 & .2 & -.1 & -.12 \\ 0 & -.7 & 0 & 0 & .4 & 0 & 0 & +.2 & 0 \\ .7 & 0 & -.6 & -.2 & 0 & -.8 & -.2 & 0 & +.12 \\ 1.5 & -.3 & .3 & -.4 & .2 & +.4 & -.5 & +.1 & -.06 \end{array}\right]$$

$$\text{SUM OF CROSS PRODUCTS} = B = \frac{1}{S} \begin{bmatrix} -1.1 & -.4 & +.36 \\ .2 & +.8 & -.02 \\ -.7 & .4 & +.20 \\ .1 & -1.0 & -.08 \\ 1.5 & .2 & -.46 \end{bmatrix} = \text{PCA SCORES}$$

MEANS OF ORIGINAL DATA (1 x 3)		EIGENVECTORS (3 x 3)		C (1 x 3)
$[\bar{X}\ \ \bar{X}_2\ \ \bar{X}_3]$		$\begin{bmatrix} .37 & -.10 & -.12 \\ -.34 & .22 & .08 \\ .28 & .39 & -.06 \end{bmatrix}$	=	$[C_1\ \ C_2\ \ C_3]$
$[24\ \ 3\ \ 6]$	X		= * =	$[9.54\ \ .6\ \ -3.0]$

$$* = [24(.37)+3(-.34)+6(.28)\quad 24(-.10)+3(.22)+6(.39)\quad 24(-.12)+3(.08)+6(-.06)]$$

$$= [8.88 - 1.02 + 1.68 \quad -2.4 + .66 + 2.34 \quad -2.88 + .24 - .36]$$

$$= [9.54 \quad .6 \quad -3.0]$$

Matrix of cross products from page 3 $= \frac{1}{S}$

$$\begin{bmatrix} .37(4) - .34(-2) + .28(-1) & -.10(-4) + .22(-2) + .39(-1) & -.12(-4) + .08(-2) - .06(-1) \\ .37(-2) \ \ldots & & \\ .37(0) \ \ldots & & \\ .37(2) \ \ldots & & \\ .37(4) \ \ldots & & \end{bmatrix}$$

LINE 1 *
LINE 2
LINE 3
LINE 4
LINE 5

$$C = \begin{bmatrix} .37(24) - .34(3) + .28(6) & -.10(24) + .22(3) + .39(6) & -.12(24) + .08(3) - .06(6) \\ E_{11}(\bar{x}_1) - E_{21}(\bar{x}_2) + E_{31}(\bar{x}_3) & -E_{12}(\bar{x}_1) + E_{22}(\bar{x}_2) + E_{32}(\bar{x}_3) & -E_{13}(\bar{x}_1) + E_{23}(\bar{x}_2) - E_{33}(\bar{x}_3) \end{bmatrix}$$

LINE 6

Matrix of cross products from page 2 =

$$\begin{bmatrix} .37(20) - .34(1) + .28(5) & -.10(20) + .22(1) + .39(5) & -.12(20) + .08(1) - .06(5) \\ E_{11}(DV)^{**} - E_{21}(DV) + E_{31}(DV) & -E_{12}(DV) + E_{22}(DV) + E_{32}(DV) & -E_{13}(DV) + E_{23}(DV) - E_{33}(DV) \\ .37(22) \ \ldots & & \\ .37(24) \ \ldots & & \\ .37(26) \ \ldots & & \\ .37(28) \ \ldots & & \end{bmatrix}$$

LINE 7
LINE 9
LINE 10
LINE 11
LINE 12

*Line 1 differs from line 7 by line 6; lines 2-5 differ from lines 9-12 by line 6.

**(DV) - (Data Value)