Ecology, 78(3), 1997, pp. 929–940

COMPOSITIONAL DATA IN COMMUNITY ECOLOGY: THE PARADIGM OR PERIL OF PROPORTIONS?

DONALD A. JACKSON

Aquatic Ecology Group, Department of Zoology, University of Toronto, Toronto, Ontario, Canada M5S 1A1

Abstract. Ecologists are often restricted to using or choose to use proportional- or percentage-type data with the view that it helps standardize for differences in variable totals among sampling units or individuals. This standardization to compositional data leads to constraints in the covariance and correlation structure that profoundly affect subsequent analysis and interpretation. This is another form of the problem related to the use of ratios in statistical analyses. Using simulated and zooplankton data I demonstrate the effect of using compositional data vs. the original data in correlation, ordination, and cluster analysis, which are common analytical methods in community ecology. Interpretations about the relatedness of various taxa or sites may reverse when using compositions relative to the unstandardized data. In addition, the selection of subcompositions (i.e., one or more variables are excluded when calculating the composition) may have profound and unpredictable consequences for the results. I examine some approaches proposed to deal with such data, e.g., centered log-ratio analysis, and recommend the use of correspondence analysis in multivariate studies to avoid the problems associated with differing solutions.

Key words: community ecology, statistics; compositional data, analysis; ipsative data; multivariate statistics; normative data; percentage, statistical analysis; proportion, statistical analysis; statistics, proportional data.

INTRODUCTION

Ecologists must often analyze data sets comprising samples varying greatly in total species abundance. In this instance species with the greatest abundance in an observation may overwhelm the analysis and subsequent relationships may simply reflect differences in absolute abundance rather than relative abundance. To compensate for this problem, ecologists often choose to convert such data to proportions, percentages, or frequencies by dividing each variable by the total for each observation prior to more detailed analysis. The rationale for this standardization is the desire to compare all samples on a similar scale, thereby ''correcting'' or removing the influence of overall species abundance. Conceptually, this approach is appealing; however, it is rarely recognized that this standardization will limit the possible range of interspecific relationships as well as the patterns among the samples. Occasionally data are converted to proportions for other reasons. For example, Gates et al. (1983) found that ordinations were easier to interpret and that greater amounts of the total variance could be explained by using proportional data. In general, the implications of this type of standardization and its consequences are not recognized by ecologists. By converting raw data to compositional data (i.e., percentages, proportions, or frequencies), changes in the covariance and correlation structure of the data matrix may lead us to conclude that particular relationships exist when such patterns are predictable artefacts of this type of standardization (Chayes 1971, Aitchison 1986).

The difficulty with using traditional or standard statistical approaches when analyzing such data is that the results obtained from an analysis of the raw data (i.e., termed the ''basis'' or ''normative data'' in the literature) and from the ''compositional'' or ''ipsative'' data lead to very different interpretations. The raw data may suggest that some variables are uncorrelated with one another, whereas the composition-based analyses may show highly correlated relationships for the same variables. As well, the reverse situation occurs frequently. The difficulty is how to reconcile these divergent results when both are available. Also, it is often the case that only the composition is available (e.g., paleolimnology, toxicology, activity budgets, feeding selectivity). Ideally, what we require is a means of analyzing the data that emphasizes relative, rather than absolute, relationships between variables and provides a similar result regardless of whether the basis or compositional data are analyzed. Such a result would permit us to compare results obtained from studies employing different enumeration methods. For example, it is common in working with zooplankton, pollen, and various other taxa that a specific number of organisms be counted, e.g., a count of 300 individuals, and then the relative proportions of each taxon in an observation be determined prior to statistical analysis. However, other researchers working with these same taxa may choose a different approach, e.g., based on total counts found in

Manuscript received 23 June 1995; revised 28 May 1996; accepted 31 May 1996; final version received 25 June 1996.

TABLE 1. Means and variances for the basis and composition for simulated data **SIM** and lake zooplankton **ZOO**.

	Basis variables for **SIM**				
Statistic	S1	S2	S3	S4	S5
Mean	30	60	60	120	120
Variance	16	16	64	64	4096
	Basis variables for **ZOO**				
	H1	H2	H3	H4	H5
Mean	83.82	33.63	266.6	95.64	43.34
Variance	7687	1211	49110	7971	1837

a specified volume. If the results are analyzed with traditional statistical approaches, then the interpretations and conclusions may depend predominantly on the method of enumeration and standardization, rather than on any inherent ecological relationships. If we can use alternative methods of analysis, as identified in this paper and the references therein, then we can compare results from different studies without concern for the constraints imposed by differences in the basis and composition, but rather focus on the ecological relationships.

The underlying principle in using proportions is to understand how one variable responds relative to another when standardized to a common scale. This has led some researchers to propose using ratios as a means of scaling variables (e.g., Mosimann and James 1979, James and McCulloch 1990). Some measure of the magnitude or size of each observation is selected (e.g., total length in morphometrics) and all variables are divided by this measure to scale the variables to a common level, and then, generally, log transformed. (Note that ratio-based analysis is not without controversy, e.g., Atchley et al. 1976, Gibson 1984, Pearson 1897, Rising and Somers 1989, Jackson and Somers 1991). This approach is used as a means of examining the pattern of ''relative'' covariation between the variables after ''standardizing'' for the magnitude or size effect.

Although standardizations are common throughout biology, I will illustrate their effects with several statistical methods commonly used by community ecologists. I use simulated data where the relationships between variables are known, as well as lake zooplankton data, to show how interpretations change dramatically depending upon whether raw or compositional data are analyzed.

METHODS

Data sets

Data set **SIM** was simulated to comprise 200 observations for each of five variables with different means and variances (see Table 1). The variables were simulated to be independent with correlations of zero (Fig. 1). For each observation (i.e., row) in this matrix, the total was calculated and each value in that row was divided by the row total. This transformed the data from the original abundances (hereafter referred to as the basis following Aitchison 1986) into proportions or, if multiplied by 100, into percentages (hereafter referred to as the composition).

The abundances of zooplankton in five size classes from 26 lakes comprised the second data set (the matrix **ZOO**; Table 1). These data are annual ice-free abundances of herbivorous zooplankton collected by the Lake Ecosystem Working Group (LEWG; Gates et al. 1983, Paloheimo and Zimmerman 1983, Zimmerman et al. 1983). These data were converted to compositional data by dividing the value for each size class by the total abundance for a given lake.

Data analysis

Statistical analyses included several methods used by ecologists and particularly community ecologists. Simple bivariate summaries based on Pearson product-moment correlation coefficients were calculated. The statistical significance of these correlations was assessed using standardized tables and randomization tests (e.g., Jackson and Somers 1991, Manly 1991). Within each matrix the pairwise correlations between variables were calculated. Values within each variable were then randomly permuted among observations, thereby destroying the original covariance structure. Using these randomized values, the correlations were recalculated. These calculations were repeated for 10 000 permuted matrices to generate a distribution of correlation coefficients when the observations are truly arranged randomly. The proportion of correlation coefficients having an equal or more extreme value is the associated degree of probability of the null hypothesis (i.e., random correlation) being true. When calculating null correlations for the compositions, the row totals for each observation in the randomized basis was determined (i.e., after the permutation step). Each value in a given row was then divided by the row total to convert it to a proportion prior to calculating the intervariable correlations.

Principal components analyses (PCA) using correlation and covariance matrices were calculated (SAS 1988). Ordination analyses are methods used by ecologists and PCA is a commonly employed and a conceptually simple method summarizing linear multivariate patterns (e.g., Legendre and Legendre 1983, Digby and Kempton 1987, Reyment 1991). A second ordination method, correspondence analysis (CA), was also used, which incorporates an implicit double centering. The resemblance measure is based on a chi-squared statistic, rather than correlation or covariance measure.

Classification is another approach often used by community ecologists. As an example of this approach, I used an agglomerative hierarchical cluster analysis (Unweighted Paired Group Method of Averaging; UPGMA) of the correlation matrices. This summarized the multivariate relationships for both the basis and the

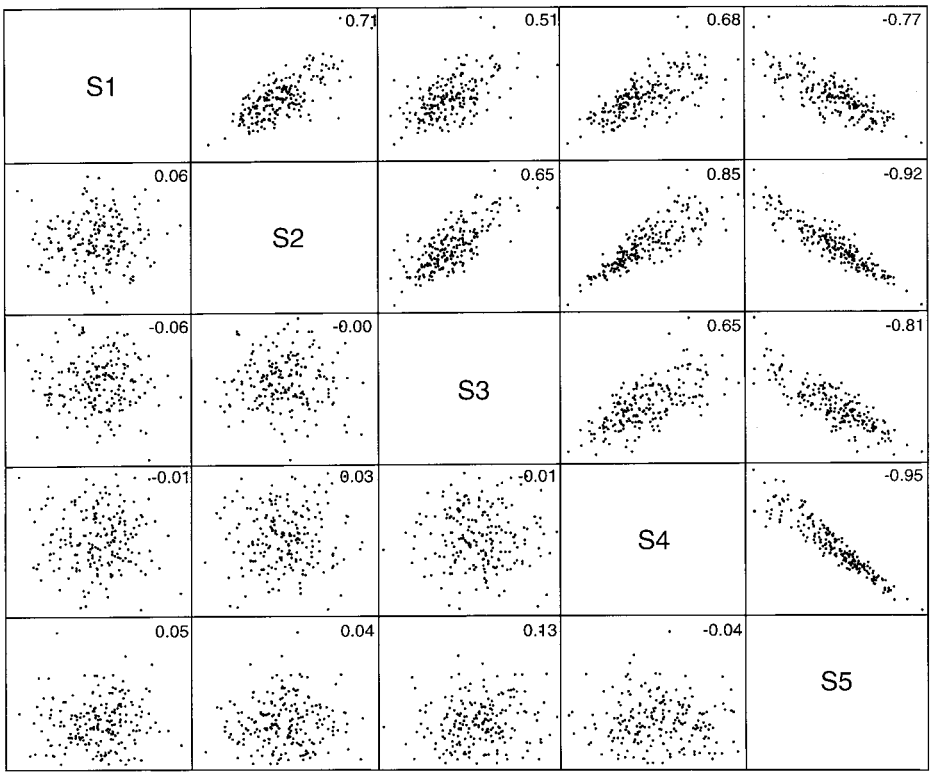

FIG. 1. Bivariate casement plots of the basis (lower triangular matrix) and composition (upper triangular matrix) for the simulated data **SIM**. The basis relationships are independently generated, and correlations approximate zero. Note the strong linear relationships in the composition arising due to the constant-sum constraint, i.e., matrix closure. S1–S5 represent variables.

composition. The two resultant dendrograms were compared visually to illustrate how the standardization changed the interpretation of the relationship among variables. An additional cluster analysis of the lakes based on the interlake Euclidean distances was completed using the basis and compositional forms of the zooplankton and simulated data. The results from the interlake comparison simulated data are not presented simply due to the large dendrogram resulting, i.e., 200 observations. The CA and cluster analyses were done using NT-SYS (Rohlf 1993).

RESULTS AND DISCUSSION

What are the implications of using proportions on covariance and correlation structure?

For the raw data, bivariate statistics revealed no evidence of statistically significant relationships (Fig. 1). However, the ''standardization'' of the basis to compositional data produced ''significant'' bivariate correlations between the variables and strong linear bivariate patterns (Fig. 1). This apparent significance is due to the constraint that each observation must sum to a constant (i.e., 1.00 or 100%). In such analyses, compositional data will be biased toward negative relationships because of the following conditions (using the notation of Aitchison [1986]). Consider a basis matrix **x** composed of D parts or variables $(x_1,\ldots x_D)$. The variance of variable i is represented as $\text{var}(x_i)$, the covariance between variables i and j as $\text{cov}(x_i, x_j)$, and the correlation as $\text{corr}(x_i, x_j)$. Using these definitions, we can generate D variances $\text{var}(x_i)$ where $i = 1, \ldots D$ and the number of covariances will be $0.5dD$ covariances $\text{cov}(x_i, x_j)$ $(d = D - 1; i = 1, \ldots d; j = i + 1, \ldots D)$, or expressed alternatively as $D(D - 1)/2$.

For any given variable in a covariance matrix of compositional data, the sum of the variances and covariances must equal zero. It then follows that the total of the variances and covariances for the entire matrix must also equal zero. This is a simple function of the constant-sum constraint. We can express the constraint on the covariance matrix as $\text{cov}(x_1x_1 + x_1, x_2 + \ldots + x_D, x_D) = 0$. Given that all the variances must be positive and the constraint above, it follows

$$\sum_i \text{var}(x_i) = \sum_{i=j} \text{cov}(x_i, x_j) = -\sum_{i.\neq j} \text{cov}(x_i, x_j).$$

We can see that given this relationship, some of the covariances must be negative. In fact, Aitchison (1986) shows that at least D of the possible $D(D - 1)/2$ covariances must be negative. This balances the D positive values representing the variances. For example, with **SIM** the sum of the variances is equal to 0.0180,

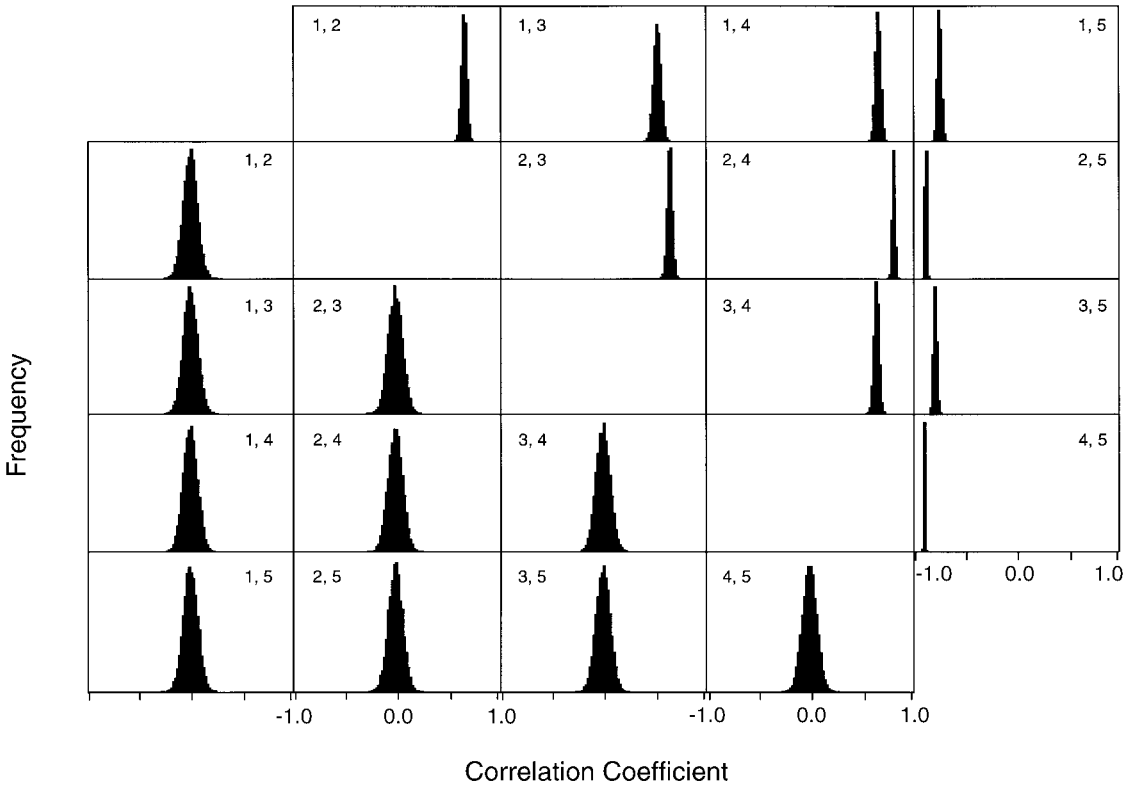

FIG. 2. Frequency distributions of the bivariate correlations for **SIM** obtained under randomization. Each plot corresponds to the correlation between two variables from the basis (lower triangular matrix) or the composition (upper triangular matrix) used in Fig. 1. The basis matrix was randomized within each column, the composition recalculated, and the correlation recalculated. Each plot is a frequency distribution of the correlations obtained from 10 000 randomized matrices.

the sum of the positive covariances is equal to 0.0081, and by definition the negative covariances must be equal to −0.0261. This condition holds for matrix **SIM** and any other composition. The implications of this condition are generally not considered. Our usual null hypothesis when assessing a correlation coefficient is that of independence between the variables, i.e., H_0: $r = 0$. Although the basis correlations approach zero, the constant-sum condition (i.e., ''matrix closure,'' although this term is no longer favored) results in highly significant correlations among variables in the compositional data (Fig. 1). For example, the basis variables S4 and S5 have a correlation of −0.036, whereas in their compositional form the correlation is −0.953. This relationship is not due to any inherent association between the original variables. It is due only to the standardization, and it, therefore, is an artefact of the standardization. Moreover, with compositional data the correlation coefficients are not free to range between −1 and 1. This constraint on the range of values is clearly illustrated in Fig. 2 and also identified by Reyment and Jöreskog (1993: 124). As a result, our standard null hypothesis regarding association is inappropriate (Chayes 1971, Jackson et al. 1989, Jackson and Somers 1991).

The randomization procedure illustrates clearly the distribution of the correlation coefficients when using compositional data (Fig. 2). Correlations based on the randomized basis data show symmetrical distributions centered on zero. This is the expected distribution of random correlations representing distributions comparable to those from classical statistical tables. However the distributions for the compositional data are no longer centered on zero. In fact the zero value, our traditional test of the null hypothesis, lies outside of the distributions for all of the bivariate relationships from the compositional data. This example is a generalization of the effect shown by Jackson et al. (1989) and Jackson and Somers (1991). Most of the compositional correlations range between 0.65 and 0.95. When variable S5 is included in a correlation, the results are correlations ranging between −0.77 and −0.95. This result is due to the relatively large mean and variance of S5 in the basis. Given these conditions in the basis, S5 has a large influence on the overall sum for each observation and subsequent calculation of each proportion, thereby resulting in the other variables being negatively correlated with S5.

The zooplankton data set also shows a similar, although less clearly defined effect. Many of the correlations change from positive to negative in direction (Fig. 3). Many of the correlations are weak in both the

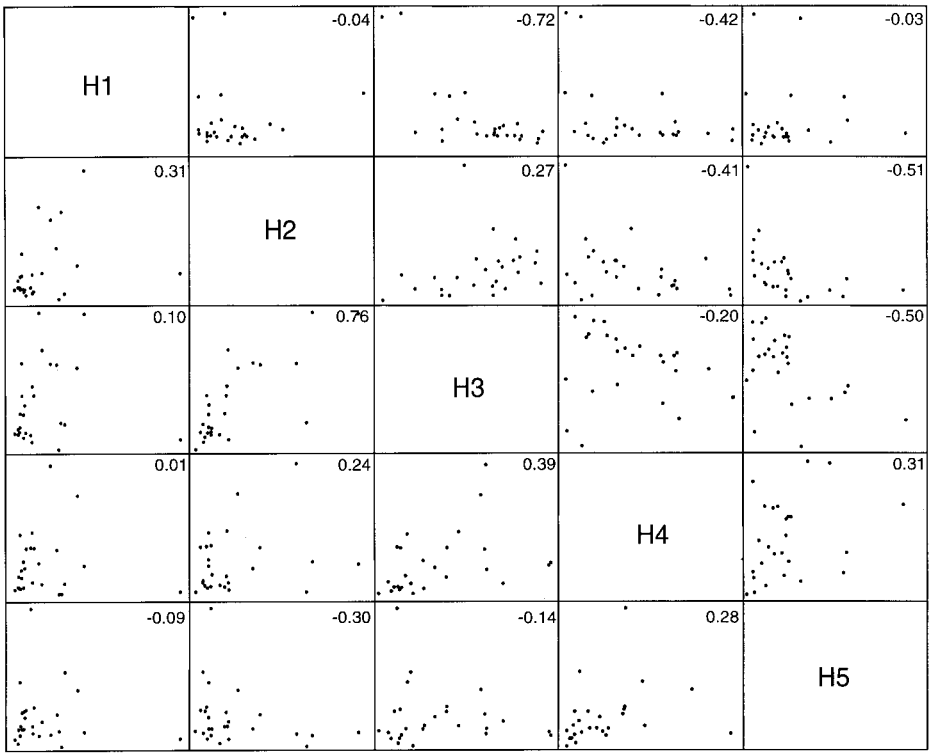

FIG. 3. Bivariate casement plots of the basis (lower triangular matrix) and composition (upper triangular matrix) for the zooplankton data **ZOO**. The basis relationships are independently generated, and correlations approximate zero. Note the strong linear relationships in the composition arising due to the constant-sum constraint, i.e., matrix closure. H1–H5 represent variables.

basis and composition, often with nonlinear patterns in the data. As such, the Pearson correlations underestimate the strength of some relationships, but are retained for consistency with the simulated data example. In several cases the correlations change from being ''nonsignificant'' with the basis to ''significant'' with the composition (e.g., H1 with H3, $r = 0.102$ and $r = -0.720$, respectively). Under the permutation procedure the derived distribution of correlations shows greater variability (Fig. 4) than that found with the simulated data (Fig. 2) due to both sample size effects and the degree of linearity between variables (i.e., greater variability in the correlation coefficients with **ZOO** due to the nonlinear nature of the bivariate relationships). Many distributions for the basis are skewed slightly, but centered on zero. Many of the composition distributions are more symmetrical, but several show a negative bias in their location with few, if any, zero values occurring (e.g., correlations between H1 and H3).

This bias in the correlation structure has long been recognized. Pearson (1897) discussed the problem of spurious correlations with the use of indices that comprise a part-whole relationship. Compositional data are a special type of ratio wherein each datum is divided by the sum of the variables for any given observation. Since Pearson's work, a better understanding of the problems of using compositional data has followed from Chayes (1960, 1971, 1983), Butler (1976, 1978, 1979*a*, *b*, 1981), Aitchison (1981, 1982, 1986) and Reyment (Reyment 1991, Reyment and Jorsekog 1993) in the statistical and geological literature. A parallel set of literature exists in the psychometrical and sociological literature (Cattell 1944, Jackson and Alwin 1980, Dunlap and Cornwall 1994), although there appears to be no recognition of the work between these fields, likely due to differences in the terminology. (The basis is referred to as normative data, whereas the composition is called ipsative data.) The magnitude of this bias depends on the number of variables used in the analysis. It is most pronounced with few variables and decreases in magnitude as the number of variables is increased.

Chayes and Kruskal (1966) proposed a test of the presence of significant correlations in compositional data. Their test was based on simulating a basis data set that produced compositional data with similar characteristics as the observed compositional data. Aitchison (1981) noted several shortcomings with their approach. For example, different basis data sets can give rise to identical compositions, and negative variances can arise in simulating basis data (Butler 1975). There is no overall test available so numerous pairwise tests must be done, and the distribution of the test statistic

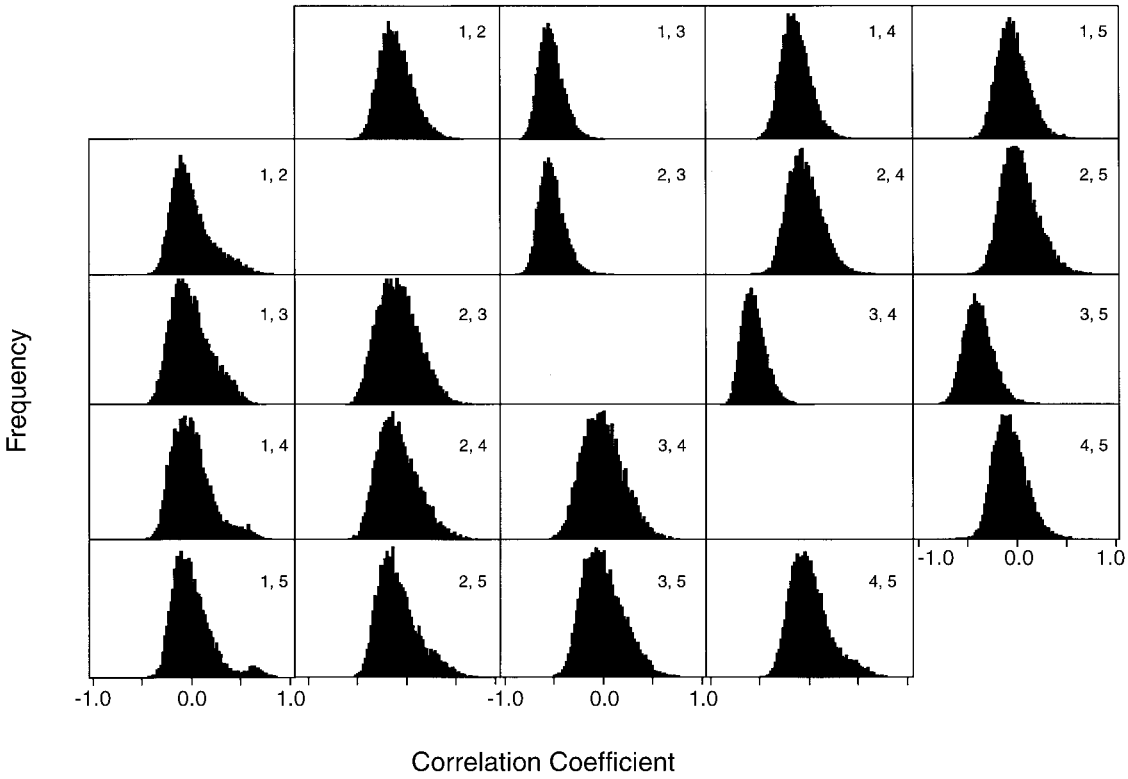

FIG. 4. Frequency distributions of the correlations for **ZOO** obtained under randomization. Each plot corresponds to the correlation between two variables from the basis (lower triangular matrix) or the composition (upper triangular matrix) used in Fig. 3. The basis matrix was randomized within each column, the composition recalculated, and the correlation recalculated. Each plot is a frequency distribution of the correlations obtained from 10 000 randomized matrices.

is unknown. As a result, Chayes and Kruskal's approach is no longer used.

Standard statistical texts (e.g., Zar 1984, Sokal and Rohlf 1995) recommend transforming proportions using an arcsine square-root transformation as a means of normalizing the distribution of points about the mean. Although this transformation corrects the problem of truncated tails in variable distributions, it will not resolve the problem of closure. For example Butler (1981) used a variety of transformations and found the transformations do not correct for the constant-sum constraint even though the intervariable or interobservation relationships may be altered.

TABLE 2. Correlation matrices for compositions from **SIM**. The upper triangle shows correlations when variable S5 is included in calculating the row totals, and the lower triangle shows correlations when variable S5 is not included in calculating the row totals, i.e., a subcomposition. Probability of $H_0 = 0$ as determined from statistical tables is shown in parentheses.

	Variables			
Variables	S1	S2	S3	S4
S1		0.711 (0.0001)	0.541 (0.0001)	0.656 (0.0001)
S2	−0.032 (0.652)		0.706 (0.0001)	0.826 (0.0001)
S3	−0.316 (0.0001)	−0.372 (0.0001)		0.668 (0.0001)
S4	−0.259 (0.0002)	−0.223 (0.0015)	−0.644 (0.0001)	

Another complication that arises during the analysis of compositional data involves the use of subcompositions, i.e., a subset of the variables (e.g., a set of four of the five variables in **SIM** would represent a subcomposition). When analyzing the correlation of two variables, one expects their relationship to remain similar regardless of whether another variable is included in the data set or not. With the zooplankton data we would expect that the correlation between two taxa (or size classes) should not change if we add additional taxa (or size classes) to the data set. Unfortunately, this is not the case with compositional data. In fact, Reyment (1989:31) states ''*as one moves from a D-part composition towards subcompositions of decreasing dimension, the correlations may fluctuate wildly in magnitude as well as in sign.*'' As an example, consider the results from **SIM** where variable S5 is excluded in determining the composition (Table 2). Variables S1 and S2 are uncorrelated ($r = -0.03$) and all other variable combinations have weak, but ''statistically significant'' negative correlations as assessed by statistical tables. The exception being variables S3–S4 with a

TABLE 3. Correlation matrices for compositions from **ZOO**. The upper triangle shows correlations when variable H3 is included in calculating the row totals, and the lower triangle shows correlations when variable H3 is not included in calculating the row totals, i.e., a subcomposition. Probability of $H_0 = 0$ as determined from statistical tables is shown in parentheses.

	Variables			
Variables	H1	H2	H4	H5
H1		−0.043 (0.835)	−0.424 (0.031)	−0.029 (0.889)
H2	−0.016 (0.939)		−0.406 (0.039)	−0.506 (0.008)
H4	−0.808 (0.0001)	−0.249 (0.220)		0.312 (0.121)
H5	−0.377 (0.058)	−0.509 (0.008)	0.033 (0.873)	

strong correlation ($r = -0.64$). Contrast this with our interpretation with results when S5 is included in calculating the proportions. In this case all correlations are strongly positive except those including correlations with S5 directly (Fig. 1). Had we chosen to delete the variable S4 when calculating the row totals and proportions, we would have found S1 and S2 to be highly correlated ($r = 0.84$), both variables S1 and S2 to be positively correlated with S3 (rather than negatively when S5 is deleted), and all other variables to have very strong negative correlations with S5. This overall correlation structure again is influenced by the constraint that the variances and covariances must sum to zero.

Matrix **ZOO** (Table 3) shows a similar, although more erratic effect. In this example variable H3 is deleted as it has the largest mean and variance as with the simulated data example. This follows the approach used with the simulated data. Some correlations remain virtually unchanged, i.e., variables H2 and H5 have correlations of −0.506 and −0.509 in the full composition and subcomposition, respectively. However other variables change more substantially, e.g., H1 and H4 have correlations of −0.424 and −0.808. Some correlations change from having significant correlations to values lacking significance at the 5% level, i.e., H2 and H4 have $r = -0.406$ and −0.249 for the full composition and subcomposition. The predictability of the resultant correlations in the subcomposition is less certain than with the simulated data **SIM**. In this field-based example we see changes in the magnitude and interpretation of the correlations between zooplankton, but not the changes in the sign of the association, although such changes exist with other field data sets. The unpredictable nature and substantial change in correlations obtained when using subcompositions make for difficult interpretations in the results. This makes traditional analyses and interpretations of compositions unreliable and confusing, although Aitchison (1986) provides suggested approaches in their analysis.

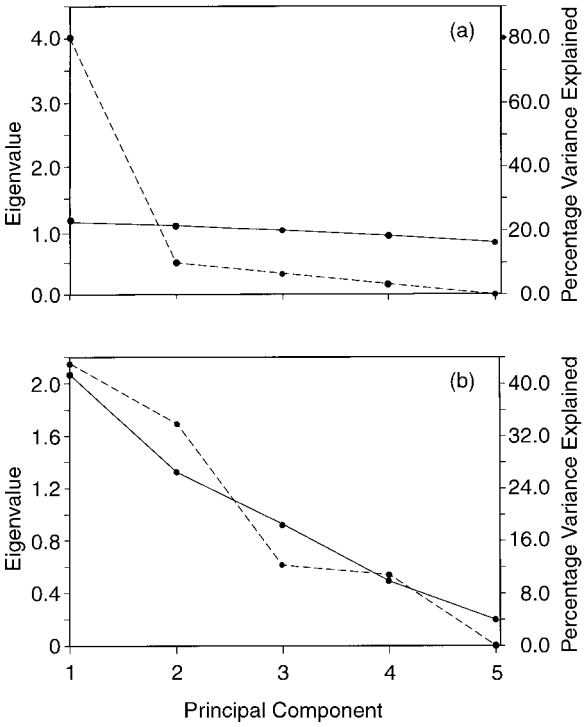

FIG. 5. Scree plots of the eigenvalues for each component from the (a) simulated data (**SIM**) and (b) herbivorous zooplankton data (**ZOO**). The solid line represents the eigenvalues from the basis data (i.e., nonstandardized), and the dashed line represents the eigenvalues from the compositional data (i.e., proportions).

Principal components analysis

As demonstrated above, strong bivariate relationships can arise as a result of converting data into compositional form. Similarly, observed correlations may be reduced or reverse in their direction. The potential for altering relationships is of even greater concern in multivariate analyses where multiple variables are considered together. Many multivariate statistical methods are used for data reduction or display as a means of ordering the observations or variables based on patterns of covariation in the data set. The implicit assumption is that the summarized covariation represents nontrivial information.

A principal components analysis (PCA) of the uncorrelated raw data **SIM** results in five eigenvalues with values near 1 and each axis accounting for ≈20% of the total variation (Fig. 5a). This result indicates that there are no nontrivial components (i.e., the variables are uncorrelated with one another based on the broken-stick model or a bootstrapped analysis; see Jackson 1993 for details). A PCA of the compositional data leads to a different interpretation (Table 4) and summary of the patterns among observations (Fig. 6). The first eigenvalue accounts for over 80% of the variation. A naïve interpretation of this result would lead to the conclusion that the PCA was a useful summary and a

TABLE 4. Eigenvector coefficients from a principal components analysis of the correlation matrix of **SIM**. Results from a PCA of the basis and the composition are presented.

	Eigenvector 1		Eigenvector 2		Eigenvector 3	
Variable	Basis	Composition	Basis	Composition	Basis	Composition
S1	0.059	0.405	0.696	0.643	−0.374	0.635
S2	0.149	0.463	0.609	0.107	0.330	−0.355
S3	0.651	0.400	−0.300	−0.755	0.279	0.470
S4	−0.203	0.464	0.185	0.045	0.819	−0.488
S5	0.714	−0.497	0.142	0.060	−0.060	0.109

single dominant component explained most of the variation in the data. The basis was composed of variables with near-zero correlations and this is illustrated by the first PCA. However, the constant-sum constraint would lead us to conclude that potentially meaningful patterns exist within the data if the underlying random nature of the raw data is unknown (i.e., nontrivial axes exist).

In general, PCA of compositional data will summarize a greater amount of the variance in the first eigenvalue than the PCA of the raw data. This is due to matrix closure and the resulting variance–covariance relationship (i.e., the resultant matrix is singular and at least the last eigenvalue of the compositional PCA must equal zero). The relative magnitude of the eigenvalues is often used to determine how many components contain nontrivial information and should, therefore, be considered meaningful. Due to considerable differences that arise between PCAs of basis and compositional data, one must be cautious about the assessment of nontrivial eigenvalues and the interpretability of components from a composition. Given the widespread acceptance of the eigenvalue magnitude as a useful guide in evaluating PCA results (Jackson 1993), it is crucial that researchers recognize this underlying bias and effect. Although I present results using a simulated data set of uncorrelated variables, researchers generally lack any knowledge of the basis or fail to examine the relationships among the raw data prior to transformation to compositional data.

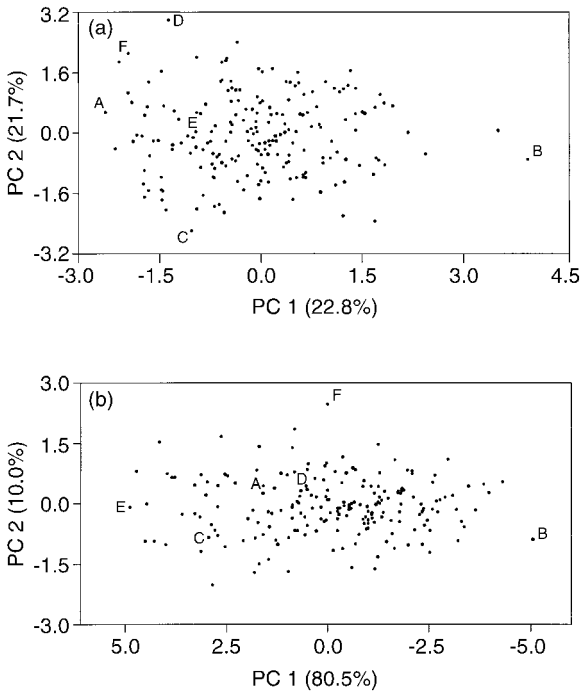

FIG. 6. Scatterplots of the first two components from a principal components analysis of **SIM** using the (a) basis and (b) composition in calculating the correlation matrix. Letters refer to the points positioned at the ends of axes 1 and 2 on the scatterplots.

Another example of this problem is illustrated with **ZOO**. Some of the variables are correlated in the basis (e.g., H2 and H3; Fig. 3) and these variables contribute most to the first eigenvalue and associated eigenvector (Table 5; Fig. 7). This relationship remains in the PCA of the compositional data set. The second component in the two analyses have similar eigenvalues, but the eigenvector coefficients differ. With the basis data, variables H4 and H5 contribute most to the second eigenvector, whereas variables H1 and H4 contribute most to this eigenvector in the compositional data. Although it is easy for researchers to rationalize such standardizations (e.g., only the patterns of relative abundance rather than total abundance are expressed in the compositional PCA), it is often unclear from subsequent analyses just how much of this result is meaningful pattern and how much of this is an artefact (i.e., due to the constant-sum constraint; see Fig. 3).

Cluster analysis

Another approach to summarizing multivariate data is cluster analysis (Legendre and Legendre 1983). For example, variables may be grouped according to their relative similarity across the observations. Such a cluster analysis of the basis correlation matrix (Fig. 8a) shows all cluster fusions occurring near zero. This is expected given the uncorrelated nature of the simulated basis data (Fig. 1). However, when the compositional data matrix is clustered, the pattern changes considerably (Fig. 8b). Variables S2 and S4 are initially clustered at a value of 0.85. Variables S1 and S3 join the S2–S4 pair at relatively high levels of positive similarity. A strong negative relationship between these

TABLE 5. Eigenvector coefficients for components 1–3 from principal component analyses of the **ZOO** basis and composition.

	Eigenvector 1		Eigenvector 2		Eigenvector 3	
Variable	Basis	Composition	Basis	Composition	Basis	Composition
H1	0.261	0.226	−0.251	−0.709	0.909	0.035
H2	0.638	−0.492	−0.125	−0.227	−0.048	0.547
H3	0.619	−0.551	−0.120	0.381	−0.237	−0.392
H4	0.320	0.333	0.650	0.525	0.022	0.664
H5	−0.196	0.541	0.696	0.155	0.339	−0.324

four variables and S5 is shown. An interpretation of this result would lead one to conclude that meaningful patterns exist such that variables S1 through S4 are strongly and positively correlated, whereas S5 is negatively related to the other variables. However, from the simulation, we recognize that these patterns are due only to the constant-sum constraint of the compositional data and not to any meaningful relationships among the variables. The cluster analysis of **ZOO** reveals differences between the basis and composition solutions. The basis shows H2 and H3 joining together with a strong positive correlation (Fig. 8c), whereas these two size classes join based on a negative correlation in the composition solution (Fig. 8d). There are considerable differences between the two dendrograms based on the group membership and level at which cluster fusion occurs.

Clustering the observations leads to a similar problem. Because we have 200 observations in **SIM,** it is impractical to show the dendrogram. However, the results differed substantially between the two dendrograms. Cluster analyses of the between-lake Euclidean

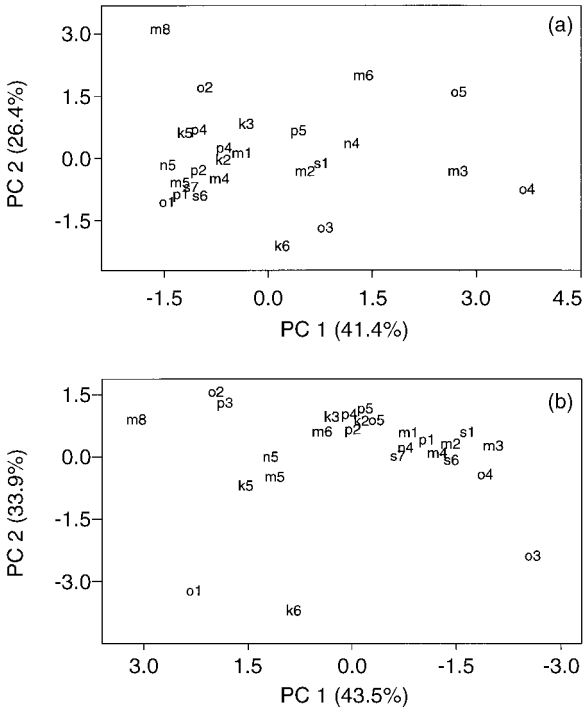

FIG. 7. Scatterplots of the first two components from a principal components analysis of **ZOO** using the (a) basis and (b) composition in calculating the correlation matrix. Letters correspond to codes for the LEWG (Lake Ecosystem Working Group) lakes (Paloheimo and Zimmerman 1993).

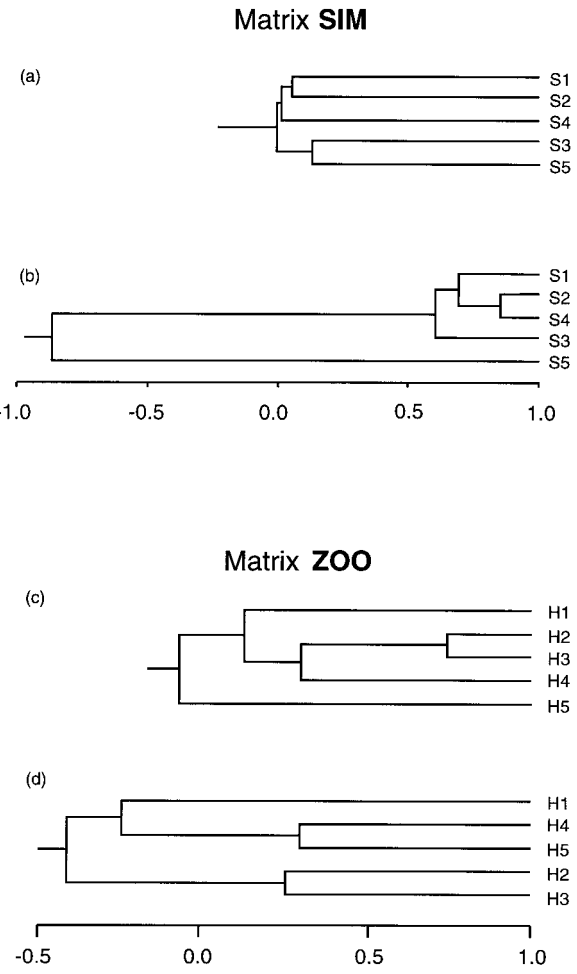

FIG. 8. UPGMA cluster analysis based on a correlation matrix of the variables (S1–S5 and H1–H5) from: (a) the basis data of the simulated data (**SIM**); (b) the compositional data of **SIM**; (c) the basis data of the zooplankton data (**ZOO**); and (d) the compositional data of **ZOO**.

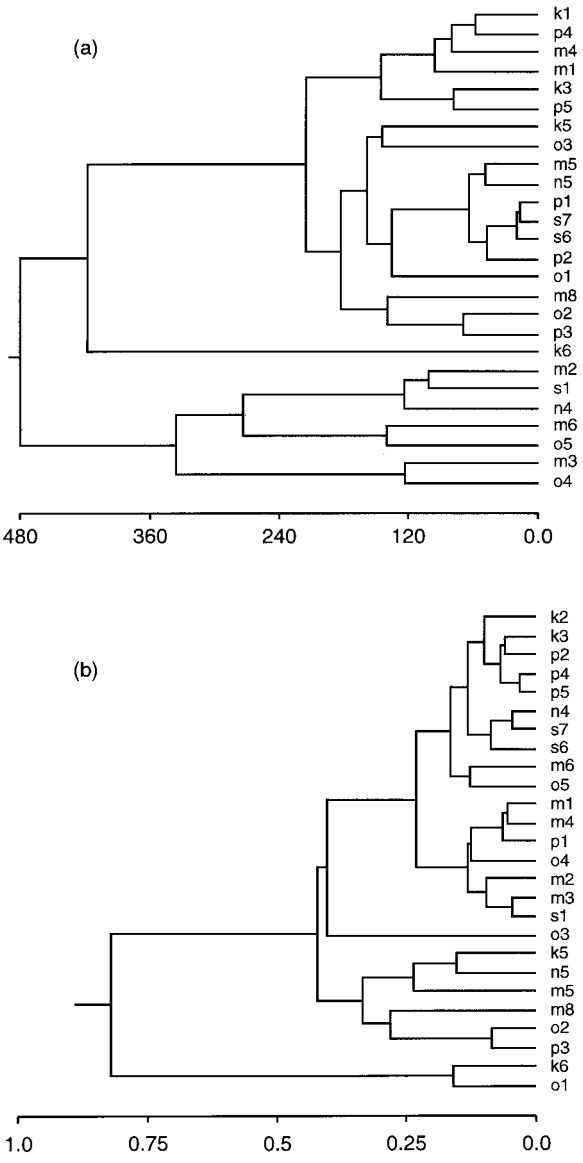

FIG. 9. Cluster analysis of the zooplankton data (**ZOO**). The letters correspond to codes for the LEWG (Lake Ecosystem Working Group) lakes (Paloheimo and Zimmerman 1983). In part (a), the lakes are clustered using UPGMA of Euclidean distances calculated from the basis data, whereas part (b) is based on the compositional data.

distances from the zooplankton basis and composition are shown in Fig. 9. These dendrograms show striking differences in both the group memberships (e.g., compare p1, s6, and s7), as well as the overall structure of the dendrograms. The dendrogram of the composition data shows a much stronger grouping in most instances (i.e., the clusters join together earlier than in the basis dendrogram).

Possible solutions to the problem

Centered log-ratios.—An approach based on centered log-ratio methods has been proposed (Aitchison 1986). Here all variables are retained in the analysis, but standardized by dividing each variable by a denominator based on a geometric composite of all variables. Specifically, the approach is based on the PCA of the covariance matrix γ where

$$\gamma_{ij} = \text{cov}\{\log[x_i/g(x)]\}, \log [x_j/g(x)]; \; i, j = 1,. . .D$$

and $g(x)$ is the geometric mean of the variables, i.e., $g(x) = \Pi\, x_i^{1/D}$.

TABLE 6. Correlation matrix of centered log-ratio variables for composition **SIM**.

Variables	Variables				
	S1	S2	S3	S4	S5
S1					
S2	0.641				
S3	0.412	0.613			
S4	0.630	0.843	0.619		
S5	−0.799	−0.906	−0.782	−0.906	

This approach has advantages in that: (1) all variables are retained and all possible pairwise comparisons are possible; (2) the pairwise relationships are identical regardless of whether the basis or compositional data are used. However, the method still has problems. (1) Although the correlations of the original simulated data, i.e., the basis, were zero, the variables in the centered log-ratio basis and composition still show strong correlations (Table 6). Because of closure and the implicit dependency of the variables on one another, as well as the division of all variables by another composite variable, the variables lack statistical independence. As a result we must be cautious in assigning statistical significance to the correlations or the use of other forms of inferential statistics (e.g., regression analysis). This is simply a modification of the problem encountered in the correlation example. In this instance, both the basis and composition have similar correlation structure, but again the null hypothesis of a correlation of zero appears to be invalid. (2) The matrix is singular and we therefore lose one eigenvalue or dimension in a principal components analysis. (3) If zero values are present in the data (e.g., no individuals were found for one or more species in one or more sampling units), then the log-ratio values are undefined. To circumvent this problem, Aitchison (1986) suggested replacing each zero value with a small numerical value. However, differences in the value chosen may lead to substantially different solutions, i.e., no unique solutions exist to alleviate the problem of zero values in log-ratio solutions. Approaches based on ranking methods have also been suggested (Bacon-Shone 1992), but he also identifies the problem of different approaches to ranking (e.g., ranking across variables, across observations, or the entire matrix) leading to different solutions.

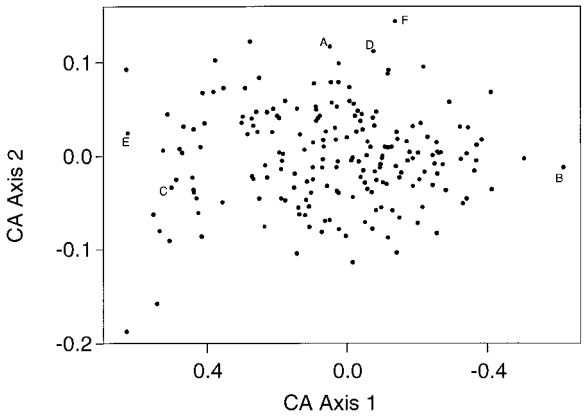

FIG. 10. Scatterplot of the first two axes from a correspondence analysis of **SIM**. The scale and position of points within a plot are identical for analyses using either the basis or composition, but the amounts of variance explained by each axis differ between data formats. The letters refer to points identified in Fig. 6.

Correspondence analysis as an alternative

Given many ecologists are trying to achieve summaries of the patterns in species-sites relationships, a solution to the problem of compositional data is correspondence analysis (CA). Since this method considers only the proportional relationships between the variables, it is unaffected by the decision to use either the basis or the compositional data. One obtains identical solutions regardless of the form of data analyzed (Fig. 10, which shows identical results for both the basis and composition). In addition the eigenvalues obtained from the CA are virtually identical for both the analysis of the basis and of the composition. As a result of these conditions, there is no ambiguity because of differences in interpretation one obtains from the basis vs. the composition. The double standardization in CA and the chi-square distance measure implicitly removes the difference between data formats. Correspondence analysis enjoys widespread acceptance in community ecology (Legendre and Legendre 1983, Digby and Kempton 1987, Jongman et al. 1987). CA has also been incorporated into the more complicated direct gradient analysis of canonical correspondence analysis (ter Braak 1987), thereby permitting direct comparisons of community and environmental data. Given the widespread development and acceptance of CA, it appears to provide a means of ordinating community relationships rather than using a log-ratio method in many instances.

Conclusion

This paper is not an exhaustive treatise on the problems of compositional data. Rather it is an introduction to the problem for ecologists and an illustration of some solutions (e.g., Aitchison 1986). In some cases we can use methods such as correspondence analysis that may be particularly well suited to avoid problems when using compositional data. We may need to consider the application of these methods more widely and formally recognize the advantages they offer over other methods (e.g., PCA). Furthermore, we need to explore the consequences of using compositions with respect to our interpretation and analyses. Consideration of those methods developed elsewhere to solve these specific problems (i.e., log-ratio methods; Aitchison 1986) is essential. This area of analysis is of considerable importance to ecologists, but given the relative unfamiliarity of biologists and statisticians to it, it is apparent that additional work on the development of methods must be done.

In many instances ecologists may have the choice of whether to analyze the basis or compositional forms of their data. In these instances we have the opportunity to assess the effect that such standardizations have on our interpretations. If similar conclusions are reached, then it is of little consequence which form of data is chosen. However in many instances we may be limited to only the composition and this case is easily recognized. In areas such as paleoecology only the proportions of pollen or other fossil organisms are available or researchers studying activity budgets or behavior may represent the amount of time for different activities as proportions. In other instances the use of compositional data may not be recognized readily. For example, the enumeration of plankton abundance is often done by counting a fixed number of organisms and representing abundances as the proportion or percentage within each taxon. This is a common approach in bioassessment methods (Plafkin et al. 1989, Novak and Bode 1992). This implicitly converts the data to a composition and imposes the constant-sum constraint (i.e., as the abundance of one taxon increases, one or more taxa must decrease in the total count). Researchers must recognize these potential cases and their consequences in examining results (e.g., relative abundance in community ecology). Criteria such as the ease of interpretation or the ability to recover greater amounts of variation are poor measures to assess whether a data standardization has been useful. We must recognize that these criteria may be biased due to the constant-sum (matrix closure) condition.

ACKNOWLEDGMENTS

I would like to thank J. Choi and K. M. Somers for their comments on previous versions of this manuscript and A. P. Zimmerman for supplying the LEWG zooplankton data set. I am grateful to R. A. Reyment and an anonymous reviewer for their comments.

LITERATURE CITED

Aitchison, J. 1981. A new approach to null correlations of proportions. Journal of Mathematical Geology **13**:175–189.

———. 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society B **44**:139–177.

———. 1986. The statistical analysis of compositional data. Chapman and Hall, New York, New York, USA.

Atchley, W. R., C. T. Gaskins, and D. Anderson. 1976. Statistical properties of ratios. I. Empirical results. Systematic Zoology **25**:137–148.

Bacon-Shone, J. 1992. Ranking methods for compositional data. Applied Statistics **41**:533–537.

Butler, J. C. 1975. Occurrence of negative open variances in ternary systems. Mathematical Geology **7**:31–45.

———. 1976. Principal component analysis using the hypothetical closed array. Journal of Mathematical Geology **8**:25–36.

———. 1978. Visual bias in R-mode dendrograms due to the effect of closure. Journal of Mathematical Geology **10**: 243–252.

———. 1979*a*. Effects of closure on the measure of similarity between samples. Journal of Mathematical Geology **11**:431–440.

———. 1979*b*. Trends in ternary petrologic variation diagrams—fact or fantasy? American Mineralogist **64**:1115–1121.

———. 1981. Effects of various transformations on the analysis of percentage data. Journal of Mathematical Geology **13**:53–68.

Cattell, R. B. 1944. Psychological measurement: normative, ipsative, interactive. Psychological Review **51**:292–303.

Chayes, F. 1960. On correlation between variables of constant sum. Journal of Geophysical Research **65**:4185–4193.

———. 1971. Ratio correlation. University of Chicago Press, Chicago, Illinois, USA.

———. 1983. Detecting nonrandom associations between proportions by tests of remaining-space variables. Journal of Mathematical Geology **15:**197–206.

Chayes, F., and W. Kruskal. 1966. An approximate statistical test for correlations between proportions. Journal of Geology **74**:692–702.

Cornwell, J. M., and W. P. Dunlap. 1994. On the questionable soundness of factoring ipsative data: a response to Saville & Wilson (1991). Journal of Occupational and Organizational Psychology **67**:89–100.

Digby, P. G. N., and R. A. Kempton. 1987. Multivariate analysis of ecological communities. Chapman and Hall, New York, New York, USA.

Dunlap, W. P., and J. M. Cornwall. 1994. Factor analysis of ipsative measures. Multivariate Behavioral Research **29**: 115–126.

Gates, M. A., A. P. Zimmerman, W. G. Sprules, and R. Kneochel. 1983. Planktonic biomass trajectories in lake ecosystems. Canadian Journal of Fisheries and Aquatic Sciences **40**:1752–1760.

Gibson, A. R. 1984. Multivariate analysis of lizard thermal behavior. Copeia 1984:267–272.

Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology **74**:2204–2214.

Jackson, D. A., H. H. Harvey, and K. M. Somers. 1989. Ratios in aquatic sciences: statistical shortcomings with mean depth and the morphoedaphic index. Canadian Journal of Fisheries and Aquatic Sciences **47**:1788–1795.

Jackson, D. A., and K. M. Somers. 1991. The spectre of spurious correlation. Oecologia **86**:147–151.

Jackson, D. J., and D. F. Alwin. 1980. The factor analysis of ipsative measures. Sociological Methodology and Research **9**:218–238.

James, F. J., and C. E. McCulloch. 1990. Data analysis and the design of experiments in ornithology. Pages 1–63 *in* R. F. Johnston, editor. Current ornithology. Volume 2. Plenum Press, New York, New York, USA.

Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren. 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen, The Netherlands.

Legendre, L., and P. Legendre. 1983. Numerical ecology. Elsevier, Amsterdam, The Netherlands.

Manly, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London, UK.

Mosimann, J. E., and F. C. James. 1979. New statistical methods for allometry with application to Florida Red-winged Blackbirds. Ecology **33**:444–459.

Novak, M. A., and R. W. Bode. 1992. Percent model affinity: a new measure of macroinvertebrate community composition. Journal of the North American Benthological Society **11**:80–85.

Paloheimo, J. E. and A. P. Zimmerman. 1983. Factors influencing phosphorus-phytoplankton relationships. Canadian Journal of Fisheries and Aquatic Sciences **40**:1804–1812.

Pearson, K. 1897. Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society **60**:489–498.

Plaftin, J. L., M. T. Barbour, K. B. Porter, S. K. Gross, and R. M. Hughes. 1989. Rapid bioassessment protocols for use in streams and rivers. U.S. Environmental Protection Agency Office of Water, **EPA/444/4-89-001.**

Reyment, R. A. 1989. Compositional data analysis. Terra Nova **1**:29–34.

———. 1991. Multidimensional paleobiology. Pergamon Press, Oxford, England.

Reyment, R. A., and K. G. Jöreskog. 1993. Applied factor analysis in the natural sciences. Cambridge University Press, New York, New York, USA.

Rising, J. D., and K. M. Somers. 1989. The measurement of overall body size in birds. Auk **106**:666–674.

Rohlf, F. J. 1993. NTSYS-pc numerical taxonomy and multivariate analysis system. Version 1.8. Exeter Software, Setauket, New York, USA.

SAS. 1988. SAS/STAT user's guide. Release 6.03 edition. SAS Institute, Cary, North Carolina, USA.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Third edition. W. H. Freeman, New York, New York, USA.

ter Braak, C. F. J. 1987. CANOCO ——— a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis. ITI-INO, Wageningen, The Netherlands.

Zar, J. H. 1984. Biostatistical analysis. Second edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Zimmerman, A. P., K. M. Noble, M. A. Gates, and J. E. Paloheimo. 1983. Physicochemical typologies of south-central Ontario lakes. Canadian Journal of Fisheries and Aquatic Sciences **40**:1788–1803.