SIMILARITY COEFFICIENTS: MEASURES OF CO-OCCURRENCE AND ASSOCIATION OR SIMPLY MEASURES OF OCCURRENCE?

DONALD A. JACKSON, KEITH M. SOMERS, AND HAROLD H. HARVEY

Department of Zoology, University of Toronto, Toronto, Ontario M5S 1A1, Canada

Submitted August 11, 1987; Revised January 22, 1988; Accepted May 1, 1988

Biologists commonly use cluster-analysis techniques to identify species assemblages and biogeographic patterns (Peters 1971; Harvey 1978, 1981; P. Legendre and Legendre 1984). In contrast, ordination techniques such as principal-coordinates analysis (PCoA; Gower 1966) and nonmetric multidimensional scaling (NMDS; Kruskal 1964*a*,*b*) have had limited use (Brown 1969; Stephenson and Williams 1971; Gauch 1982), even though ordination-based methods have shown promise (Hughes 1973; Fasham 1977; Gauch et al. 1977; Clymo 1980; Del Moral 1980; Kenkel and Booth 1987).

The widespread use of clustering procedures has revealed a number of problems with cluster analysis (see, e.g., Williams et al. 1971; Everitt 1979). Perhaps foremost among these problems is that the objective nature of cluster analysis is compromised by the subjective choices of clustering method and measures of similarity, since both the method and the measure affect the analytical outcome (Orlóci 1978; L. Legendre and Legendre 1983; Pielou 1984). Additionally, cluster-analysis techniques produce clusters even when they do not exist (Orlóci 1967; Jain et al. 1986); tied values in the similarity matrix result in a number of different dendrograms (Hart 1983); and some methods produce clusters of nonconformists or rare species (Williams et al. 1971; Noy-Meir 1973*b*; Clifford and Stephenson 1975).

Alternatively, PCoA ordinates individuals by employing an eigenanalysis of a matrix of distances between individuals (for details, see L. Legendre and Legendre 1983; Pielou 1984). Principal-coordinates analysis was originally proposed for use with Euclidean measures (Gower 1966), but Cailliez and Pagès (1976) and Sibson (1979) have suggested that this restriction may be relaxed since non-Euclidean measures produce acceptable results. Nonmetric multidimensional scaling also ordinates individuals according to an initial matrix of similarities or distances, but the linear constraints of PCoA are relaxed such that monotonic rather than strict linear arrangements of the points are reproduced (for details, see L. Legendre and Legendre 1983; Kenkel and Orlóci 1986). Consequently, the more robust NMDS permits the use of both metric and nonmetric similarity or distance measures (for a discussion of Euclidean, metric, and nonmetric measures, see Gower and Legendre 1986).

Am. Nat. 1989. Vol. 133, pp. 436–453.

A precursor to clustering and ordination is the construction of an intermediate matrix measuring similarity (or distance) between samples or between species (i.e., a Q or R mode of analysis, respectively). This step is usually implicit in the more common ordination techniques like principal-components analysis and correspondence analysis. However, in cluster analysis, PCoA, and NMDS there are many similarity coefficients and, hence, many resemblance matrices to choose from (see Hubálek 1982). In biogeographic and community studies, this choice is often limited to qualitative coefficients because estimates of relative abundance are highly variable. and quantitative sampling is thus prohibitively expensive (Lamont and Grant 1979). In studies employing quantitative data, methodological comparisons are common (Hughes 1973; Austin 1976; Fasham 1977; Kenkel and Orlóci 1986), but for binary (presence/absence) data, the choice of qualitative coefficients and the resulting analyses are frequently debated (Farris 1979; Janowitz 1979, 1980; see also Hubálek 1982; Kenkel and Booth 1987). Occasionally, dendrograms based on different coefficients are presented for the same data, depending on whether samples or species are clustered (Sepkowski and Rex 1974).

Similarity coefficients are often classified into two categories (see, e.g., Sneath and Sokal 1973; Clifford and Stephenson 1975). One group is composed of measures of co-occurrence that range from 0 to 1.0. These coefficients can be recognized by their numerator, which usually consists of a or $a + d$ (see below and table 2). The second group consists of coefficients of association, which generally range from -1.0 to 1.0. Numerators in coefficients of association generally contain $ad - bc$, where a, b, c, and d are derived from the following two-by-two contingency table (+ represents species presence, − represents absence, and $a + b + c + d = N$, the total number of samples):

$$\begin{array}{cccc} & & \multicolumn{2}{c}{\text{Species Y}} \\ & & + & - \\ \text{Species X} & + & a & b \\ & - & c & d \end{array} .$$

Studies comparing similarity coefficients are common (e.g., Cheetham and Hazel 1969; Baroni-Urbani and Buser 1976; Simberloff and Connor 1979; Hubálek 1982; Gower and Legendre 1986), but each study has generated different conclusions, prompting a general acceptance that the "behavior" of similarity coefficients is data-specific (i.e., dependent on the relative frequency of ones and zeros; see Janowitz 1980; Hubálek 1982). As a result, the choice of a similarity coefficient is largely subjective and often based on tradition or on a posteriori criteria such as the "interpretability" of the results. As Gordon suggested, "human ingenuity is quite capable of providing a *post hoc* justification of dubious classifications" (1987, p. 127).

Since the results of cluster analysis and ordination may depend on the choice of similarity coefficient (see Orlóci 1978; L. Legendre and Legendre 1983; Pielou 1984), we need to understand the behavior of different types of coefficients. The implications of choosing between clustering methods are well established (Clifford and Stephenson 1975; Orlóci 1978; Pielou 1984), but Hubálek (1982), Gower and Legendre (1986), and Kenkel and Orlóci (1986) emphasized that a thorough understanding of the consequences of choosing a particular similarity coefficient is lacking and that comparative studies are necessary.

To examine the implications of choosing different similarity coefficients, we compare the results from cluster analysis, PCoA, and NMDS using eight common similarity coefficients. We show that many of the coefficients produce comparable results irrespective of the analytical technique but that the choice of similarity coefficient greatly affects the analysis. We believe that conflicting results from co-occurrence and association coefficients reflect inherent mathematical transformations that have long been recognized in multivariate analyses of continuous quantitative data.

METHODS

Data Collection

We intensively surveyed 52 lakes in the watersheds of the Black and Hollow rivers of south-central Ontario to determine fish species composition (Jackson 1988). Species were recorded as present or absent only. Lakes were sampled with experimental gill nets, fine- and coarse-mesh trap nets, plastic traps, baited minnow traps, and seine nets (see Harvey 1978, 1981; Somers and Harvey 1984; Jackson 1988). A total of 31 species was caught. Only species occurring in more than one lake were used in the analysis ($N = 25$; table 1).

Statistical Analysis

The 52-by-25 lake-by-species data matrix was used to compare species interrelationships revealed by R-mode cluster analysis and two methods of ordination. Eight different similarity coefficients were used: Jaccard, Sørensen-Dice (originally described in Czekanowski 1913 but more frequently called the Sørensen or Dice coefficient), Russell and Rao, Simple Matching, Rogers-Tanimoto, Ochiai, Yule, and Phi (see table 2). The first six coefficients are co-occurrence measures, whereas the last two are measures of association. Similarities derived from these coefficients were also transformed to distance measures by taking the square root of the complement (i.e., $(1 - S)^{1/2}$). All the resulting distance measures have Euclidean properties with the exception of Yule's coefficient, which is nonmetric (table 2; see Gower and Legendre 1986). Euclidean coefficients are metric, although not all metric coefficients have Euclidean properties.

All multivariate analyses were completed using NT-SYS numerical-taxonomy package (Rohlf et al. 1982). Dendrograms for all eight similarity coefficients were constructed using the unweighted paired-group method of averaging (UPGMA; Sneath and Sokal 1973). Dendrograms were subjectively compared using visual inspection and then contrasted with cophenetic correlation coefficients (Sneath and Sokal 1973) and with consensus trees using the *CI*(*C*) index (for details, see Rohlf 1982). Both the cophenetic correlation and the consensus index provide

TABLE 1

COMMON NAMES, SCIENTIFIC NAMES, AND THE FREQUENCY OF OCCURRENCE OF FISH SPECIES

Number of Lakes	Family	Species	Common Name
47	Centrarchidae	*Lepomis gibbosus*	pumpkinseed
43	Percidae	*Perca flavescens*	yellow perch
41	Catostomidae	*Catostomus commersoni*	white sucker
39	Cyprinidae	*Semotilus atromaculatus*	creek chub
35	Ictaluridae	*Ictalurus nebulosus*	brown bullhead
23	Cyprinidae	*Phoxinus eos*	northern redbelly dace
23	Cyprinidae	*Notemigonus crysoleucas*	golden shiner
19	Salmonidae	*Salvelinus fontinalis*	brook trout
16	Cyprinidae	*Notropis cornutus*	common shiner
16	Centrarchidae	*Micropterus dolomieui*	smallmouth bass
15	Cyprinidae	*Pimephales notatus*	bluntnose minnow
14	Cyprinidae	*Semotilus margarita*	pearl dace
12	Cyprinidae	*Phoxinus neogaeus*	finescale dace
11	Centrarchidae	*Micropterus salmoides*	largemouth bass
11	Cyprinidae	*Notropis heterolepis*	blacknose shiner
9	Gadidae	*Lota lota*	burbot
9	Cyprinidae	*Pimephales promelas*	fathead minnow
8	Salmonidae	*Salvelinus namaycush*	lake trout
6	Gasterosteidae	*Culaea inconstans*	brook stickleback
4	Percidae	*Etheostoma exile*	Iowa darter
4	Cyprinidae	*Rhinichthys atratulus*	blacknose dace
3	Cyprinidae	*Couesius plumbeus*	lake chub
3	Centrarchidae	*Ambloplites rupestris*	rock bass
3	Salmonidae	*Coregonus artedii*	cisco
2	Cyprinidae	*Semotilus corporalis*	fallfish

relative estimates of dendrogram similarity. The cophenetic correlation incorporates information associated with cluster membership and relative hierarchical position of each subcluster. In contrast, the consensus index estimates relative dendrogram congruence solely on the basis of cluster membership. That is, consensus indexes measure concordance between dendrograms as a function of the number of subsets (i.e., species found jointly in clusters in both dendrograms) relative to a measure of the total number of possible subsets (Rohlf 1982).

Both similarity and distance coefficients were used in PCoA (for details, see Gower 1966). However, only distance measures were compared in NMDS (see Kruskal 1964*a*,*b*), since distances are monotonic transformations of the similarities and provide identical results. A random initial configuration was chosen for the NMDS to minimize the possibility of local minima imposed by Euclidean approximations (see Kenkel and Orlóci 1986). Duplicate analyses using random configurations were also completed to further reduce the probability of encountering local minima in the NMDS. If the replicated solutions differed substantially, several additional runs were completed to resolve the differences. Only three-dimensional solutions were considered in this study, since these are easily compared visually (Shepard 1974).

Scores on the first three axes of all ordinations were compared with Spearman

TABLE 2

Algorithms and Metric Properties of the Eight Similarity Coefficients

		Metric Properties†		
Coefficient	Algorithm* (S)	$(1 - S)$	$(1 - S)^{1/2}$	Source
Jaccard	$\dfrac{a}{a + b + c}$	M	E	Jaccard 1901
Sørensen-Dice	$\dfrac{2a}{2a + b + c}$	N	E	Czekanowski 1913; Dice 1945; Sørensen 1948
Russell and Rao	$\dfrac{a}{a + b + c + d}$	M	E	Russell & Rao 1940
Simple Matching	$\dfrac{a + d}{a + b + c + d}$	M	E	Sokal & Michener 1958
Rogers-Tanimoto	$\dfrac{a + d}{a + 2b + 2c + d}$	M	E	Rogers & Tanimoto 1960
Ochiai	$\dfrac{a}{[(a + b)(a + c)]^{1/2}}$	N	E	Ochiai 1957
Phi	$\dfrac{ad - bc}{[(a + b)(a + c)(b + d)(c + d)]^{1/2}}$	N	E	Yule 1912
Yule	$\dfrac{ad - bc}{ad + bc}$	N	N	Yule 1900

* S is the similarity measure; see the text for explanation of symbols.

† E, Euclidean coefficient; M, metric coefficient; N, nonmetric coefficient.

rank correlations. Because rank-order changes, rather than absolute changes, were considered most important in the resulting ordinations, only rank correlations are discussed (this is consistent with the NMDS methodology). Levels of statistical significance are not given because the analyses are derived from a single initial data matrix and therefore lack independence.

RESULTS AND DISCUSSION

Cluster Analysis

The dendrograms provide little evidence of strong group structure (fig. 1). Redundancy among the coefficients is readily apparent from the nearly identical results obtained using the Jaccard and Sørensen-Dice coefficients or using the Simple Matching and Rogers-Tanimoto coefficients (table 3).

The coefficients of Jaccard, Sørensen-Dice, and Russell and Rao all initiate cluster formation among species having the greatest frequency of occurrence (i.e., the most frequently occurring species). A "chaining" of species also occurs, with less frequent species being incorporated into existing clusters. Simple Matching and Rogers-Tanimoto coefficients initiate clusters from both rare and ubiquitous species because of the inclusion of joint absences in the numerators of these coefficients.

The dendrogram based on Ochiai's coefficient also appears to summarize frequency of occurrence (fig. 1). This pattern is weaker than with the other coefficients of co-occurrence, and the chaining effect is also less apparent. The dendrogram based on Ochiai's coefficient resembles the Jaccard and Sørensen-Dice dendrograms, indicating strong similarities in cluster structure (table 3). There is little resemblance in the subcluster structure among the remaining dendrograms (i.e., using the consensus index), but including hierarchical information in the cophenetic correlation indicates that the Simple Matching and Rogers-Tanimoto dendrograms are unique.

Yule's coefficient produces tight clusters with many species pairs having similarities of 1.00 (fig. 1). This occurs when pairs of species display no mismatches (i.e., b or $c = 0$; see formula in table 2) and the coefficient assumes a value of 1.00 (for details, see Michael 1920). This problem distorts some species relationships in the resulting dendrogram. In contrast, the Phi coefficient is the only measure not markedly influenced by frequency of occurrence and not suffering the excess number of 1.00 values found in Yule's coefficient (fig. 1). However, the dendrogram based on the Phi coefficient undoubtedly contains some distortions, since zero totals in a, b, c, or d bias the coefficient (Michael 1920; Baroni-Urbani and Buser 1976).

The similar patterns in some dendrograms are not surprising because generalizations about the properties of several coefficients are possible (see Hubálek 1982; Gower and Legendre 1986). The Jaccard and Sørensen-Dice coefficients are equivalent except that double weighting is given to co-occurrences in the Sørensen-Dice coefficient (see table 2). The Simple Matching and Rogers-Tanimoto coefficients include joint absences (i.e., d) but differ in that Rogers-Tanimoto

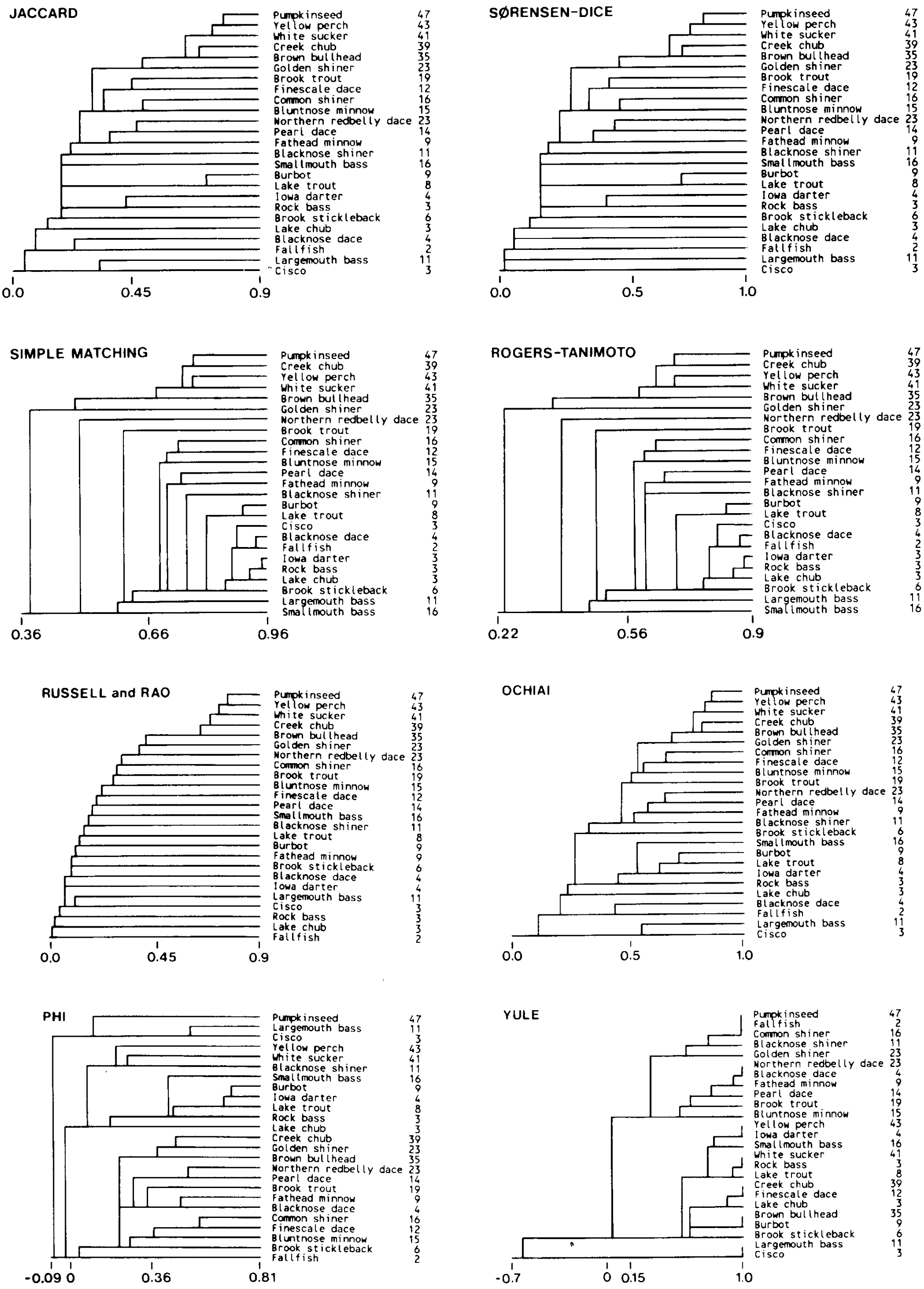

FIG. 1.—Dendrograms constructed using the unweighted paired-group method of averaging (UPGMA) and based on eight similarity coefficients. The number of lakes in which each species was found follows the species name.

TABLE 3

COMPARISON OF EIGHT DENDROGRAMS USING LOOSE CONSENSUS TREES (UPPER TRIANGLE) AND COPHENETIC CORRELATION COEFFICIENT (LOWER TRIANGLE)

Coefficient	J	SD	RR	SM	RT	O	P	Y
Jaccard		0.913	0.217	0.174	0.174	0.826	0.174	0.087
Sørensen-Dice	0.983		0.217	0.217	0.217	0.739	0.087	0.130
Russell and Rao	0.881	0.837		0.130	0.130	0.217	0.044	0.044
Simple Matching	−0.048	−0.089	−0.118		1.000	0.261	0.000	0.087
Rogers-Tanimoto	−0.060	−0.099	−0.136	0.994		0.261	0.000	0.087
Ochiai	0.971	0.988	0.815	−0.108	−0.117		0.261	0.087
Phi	0.361	0.409	0.080	0.133	0.131	0.452		0.087
Yule	0.368	0.427	0.184	−0.084	−0.077	0.476	0.523	

NOTE.—For details regarding consensus trees, see Rohlf et al. (1982). The values are the $CI(C)$ measure. J, Jaccard; SD, Sørensen-Dice; RR, Russell and Rao; SM, Simple Matching; RT, Rogers-Tanimoto; O, Ochiai; P, Phi; Y, Yule.

gives double weight to mismatches (i.e., *b* and *c*). This inclusion of joint absences provides equal importance to species presences and absences, and rare species are therefore as important as ubiquitous species in cluster formation (see, e.g., fig. 1). Conversely, coefficients including only co-occurrences (i.e., *a*) in the numerator initiate clusters with ubiquitous species. This is readily seen in the dendrogram based on Russell and Rao's coefficient, as well as in the Jaccard, Sørensen-Dice, and Ochiai dendrograms.

The Phi and Yule coefficients are based on measures of statistical association (i.e., χ^2 measures). Maximal values for these types of coefficients occur when two species are found in approximately 50% of the samples (Cole 1949; Hurlbert 1969; Fienberg and Gilbert 1970). As a result, neither ubiquitous nor rare species are favored in the Phi-based dendrogram (fig. 1). But since the analysis based on Yule's coefficient is distorted whenever the occurrence of one species is nested within a second species (i.e., b or $c = 0$; see Michael 1920), the dendrogram using Yule's coefficient tends to form pairs based on a common or ubiquitous species and a rare species. Consequently, the associations depicted by the dendrogram using Yule's coefficient contrast with all other patterns (table 3).

Principal-Coordinates Analysis

The amount of variation explained in the first three principal-coordinates-analysis (PCoA) axes ranged from 16.4% for Russell and Rao's to 58.7% for the Simple Matching coefficient. This percentage was determined using the method of Cailliez and Pagès (1976), which corrects for the occurrence of negative eigenvalues. Rank correlations of the scores obtained from analyses of both similarity and distance measures of each coefficient showed that the PCoA solutions were identical in almost all cases, even though the similarity coefficients were not metric. Since the distance matrix **D** was calculated as $(1 - S)^{1/2}$, the ordination distances based on the PCoA of the similarity matrix **S** should equal $(2)^{1/2}$ times the ordination distances derived from the PCoA of the matrix **D** (see Gower 1971). Because PCoA axes derived from similarity and distance matrices show marked correlations, only the results for the distance matrices are discussed below.

High rank correlations between scores (i.e., the position of each species) on the first axis of the PCoA and the number of lakes in which a given species occurred were obtained for all coefficients except Ochiai, Phi, and Yule (table 4). Consequently, for five of the six co-occurrence coefficients, the first axis in the PCoA appears to reflect a general "size" factor associated with the frequency of occurrence. Monte Carlo simulations have previously identified this dependence in Jaccard's coefficient (Rice and Belland 1982), but this reoccurring pattern in PCoA is similar to the general size effect often found in principal-components analysis (Jolicoeur and Mosimann 1960; Somers 1986). The magnitude of the size influence varies among coefficients (table 4, PCoA-axis 1), but the first PCoA axes derived from analyses using association coefficients are relatively free of the size influence. This effect is surprising since patterns of species association derived from measures of co-occurrence (except Ochiai) may actually reflect the frequency of occurrence and not ecological interactions. (Note also that by comparing the dendrograms [fig. 1] and the first axis of the PCoA [table 4; fig. 2], it is

TABLE 4

SPEARMAN RANK CORRELATIONS BETWEEN "SIZE" AND ORDINATION SCORES FROM PRINCIPAL-COORDINATES ANALYSIS (PCoA) AND NONMETRIC MULTIDIMENSIONAL SCALING (NMDS)

	PCoA-Axis Number			NMDS-Axis Number		
Coefficient*	1	2	3	1	2	3
Jaccard	0.942	0.281	0.117	0.575	0.353	0.136
Sørensen-Dice	0.793	0.536	0.153	0.575	0.353	0.136
Russell and Rao	0.913	0.088	0.132	0.250	0.192	0.069
Simple Matching	0.996	0.003	0.044	0.903	0.727	0.826
Rogers-Tanimoto	0.996	0.030	0.022	0.903	0.727	0.826
Ochiai	0.329	0.060	0.877	0.243	0.180	0.162
Phi	0.006	0.038	0.640	0.474	0.167	0.014
Yule	0.076	0.126	0.069	0.043	0.077	0.178

* Distance forms of coefficients only; axes based on similarities provide nearly identical results.

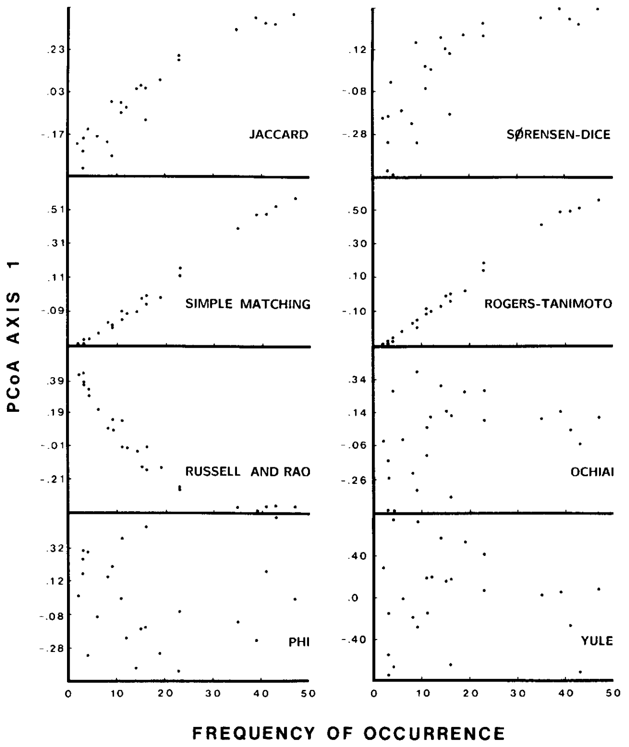

FIG. 2.—Scattergrams of species' frequency of occurrence versus axis 1 from principal-coordinates analysis for eight coefficients. Corresponding Spearman rank correlations are found in table 4.

evident that the species' rank order in the dendrograms and PCoA-axis 1 is similar; i.e., PCoA-axis 1 approximates the dendrogram result.)

Scatter plots of the first PCoA axes with the species' frequency of occurrence illustrate how the various coefficients summarize size variation (table 4; fig. 2). Both the Jaccard and Sørensen-Dice measures produce a first axis exhibiting an asymptotic relationship with frequency of occurrence. The first PCoA axis using Russell and Rao's coefficient is also asymptotic with size. However, the first axes produced from Simple Matching and Rogers-Tanimoto coefficients are linear functions of the frequency of occurrence with correlations approaching one. The inclusion of joint absences (*d*) in both the numerator and denominator of these co-occurrence coefficients appears to produce a first axis that is linear rather than asymptotic with frequency of occurrence. In contrast, first axes based on Ochiai, Phi, and Yule coefficients are not correlated with frequency of occurrence, but the size effect emerges in the third axis associated with the Ochiai and Phi coefficients.

Nonmetric Multidimensional Scaling

The nonmetric ordination is similarly affected by frequency of occurrence (table 4). The first axis in each ordination based on co-occurrence coefficients was correlated with frequency of occurrence (with the exception of the Russell and Rao and the Ochiai coefficients). Again, results from the Phi and Yule coefficients were less influenced by size effects. The arrangement of species on the nonmetric-multidimensional-scaling (NMDS) axes was similar to that found in the PCoA, but the sequence of the axes often differed (table 5). The first PCoA axis was correlated with either the first NMDS axis or with one or more subsequent axes. Since NMDS does not maximize the variance explained by each successive axis, the order of the axes based on random initial configurations may not correspond to that in PCoA. Although the size effect is often weaker in NMDS than in PCoA, the NMDS solutions retain this influence (table 4).

Comparisons among Coefficients

Correlations among axes from the various coefficients indicated a high degree of redundancy among analyses using different coefficients (table 6). Correlations between axes of the Jaccard and Sørensen-Dice ordinations approached one for both PCoA and NMDS solutions. Results from analyses using Simple Matching and Rogers-Tanimoto coefficients were also nearly identical. The first PCoA axis of each of these four coefficients and the Russell and Rao coefficient contained nearly identical information, since the relative order of the species remained the same (i.e., similar size effects). Results of ordinations using the Jaccard and Sørensen-Dice coefficients were virtually identical in all three dimensions, as were the results from the Simple Matching and Rogers-Tanimoto coefficients.

The first PCoA axis obtained from each coefficient of association was not correlated with the first PCoA axis derived from any co-occurrence measure except the Ochiai coefficient. Surprisingly, the first axis from both association coefficients was correlated with the second axis of the measures of co-occurrence (table 6). The second association-based PCoA axes were correlated with the third

TABLE 5

Spearman Rank Correlations between Ordination Scores from Principal-Coordinates Analysis (PCoA) and Nonmetric Multidimensional Scaling (NMDS)

		NMDS-Axis Number		
Coefficient	PCoA-Axis Number	1	2	3
Jaccard	1	0.518	0.594	0.155
	2	0.410	0.724	0.262
	3	0.383	0.028	0.862
Sørensen-Dice	1	0.412	0.816	0.083
	2	0.571	0.486	0.411
	3	0.345	0.163	0.871
Russell and Rao	1	0.417	0.383	0.182
	2	0.254	0.850	0.351
	3	0.412	0.197	0.539
Simple Matching	1	0.910	0.712	0.849
	2	0.364	0.473	0.397
	3	0.046	0.307	0.197
Rogers-Tanimoto	1	0.895	0.726	0.854
	2	0.402	0.430	0.350
	3	0.022	0.314	0.220
Ochiai	1	0.079	0.956	0.035
	2	0.483	0.034	0.734
	3	0.470	0.134	0.114
Phi	1	0.903	0.509	0.007
	2	0.244	0.198	0.805
	3	0.677	0.316	0.128
Yule	1	0.298	0.962	0.001
	2	0.814	0.033	0.728
	3	0.072	0.025	0.582

axes of the other coefficients, and the size influence reappeared in the third Phi axis (table 4). Although the information in the first axes from coefficients of association is independent of the size effect, the first axes were correlated with the second axes from analyses using co-occurrence coefficients. This suggests that the first axes from ordinations using co-occurrence coefficients are measures of size, whereas the first axes from ordinations of association coefficients (including Ochiai's coefficient) summarize information about the "shape" of the species assemblages (i.e., expressing interspecific associations independent of the frequency of occurrence). This shape information is expressed on second and subsequent PCoA axes of the co-occurrence coefficients. Such size and shape patterns are less evident in the NMDS ordinations (e.g., see table 4), probably because of the monotonic constraints and scaling features of the NMDS procedure.

Implication of Size Dependence

The size dependence of the first PCoA axis derived from co-occurrence coefficients has important implications in both cluster analysis and ordination. Dendrograms produced by the unweighted paired-group method of averaging show a predominant size effect as well as redundancy among the coefficients (fig.

TABLE 6

Spearman Rank Correlations between Ordination Axes Using Different Coefficients in Principal-Coordinates Analysis (upper triangle) and in Nonmetric Multidimensional Scaling (lower triangle)

Coefficient	J1	J2	J3	SD1	SD2	SD3	RR1	RR2	RR3	SM1	SM2	SM3	RT1	RT2	RT3	O1	O2	O3	P1	P2	P3	Y1	Y2	Y3
J1				98	—	—	94	—	—	93	—	—	94	—	—	—	62	—	—	73	—	—	—	—
J2				—	91	—	—	—	—	—	83	—	—	92	—	81	—	—	88	—	—	90	—	—
J3				—	—	92	—	—	—	—	—	71	—	—	75	—	88	—	—	79	—	—	73	—
SD1	100	—	—				91	—	—	88	—	—	—	—	—	70	—	63	—	—	—	—	—	—
SD2	—	100	—				—	78	—	—	72	—	—	78	—	—	—	75	75	—	—	75	—	—
SD3	—	—	100				—	—	—	—	—	85	—	—	87	—	91	—	—	89	—	—	77	—
RR1	—	—	—	—	—	—				98	—	—	98	—	—	—	—	81	—	—	63	—	—	—
RR2	—	—	—	—	—	—				—	75	—	—	78	—	83	—	—	82	—	—	91	—	—
RR3	—	—	—	—	—	—				—	—	—	—	—	—	—	81	—	—	—	—	—	88	—
SM1	74	—	—	74	—	—	—	—	—				100	—	—	—	—	88	—	—	66	—	—	—
SM2	—	89	—	—	89	—	—	—	—				—	99	—	86	—	—	97	—	—	82	—	—
SM3	—	—	—	—	—	—	—	—	—				—	—	96	—	70	—	—	92	—	—	61	—
RT1	74	—	—	74	—	—	—	—	—	100	—	—				—	—	86	—	—	64	—	—	—
RT2	—	—	—	—	—	—	—	—	—	—	100	—				—	86	—	97	—	—	84	—	—
RT3	—	—	—	—	—	—	—	—	—	—	—	100				—	70	—	—	85	—	—	54	—
O1	—	—	—	—	—	—	87	—	—	—	—	—	—	—	—				92	—	—	88	—	—
O2	—	93	—	—	93	—	—	95	—	—	—	—	—	—	—				—	84	—	—	89	—
O3	—	—	65	—	—	—	—	—	66	—	—	—	—	—	—				—	—	69	—	—	—
P1	81	—	—	81	—	—	—	—	—	74	—	—	—	—	—	—	—	—				88	—	—
P2	—	94	—	—	94	—	—	89	—	—	—	—	—	—	—	—	—	—				—	82	—
P3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	91				—	—	—
Y1	—	—	65	—	—	65	—	—	64	—	—	—	—	—	—	—	—	—	—	—	65			
Y2	—	87	—	—	87	—	—	95	—	—	—	—	—	—	—	—	94	—	—	93	—			
Y3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	61	—	62			

Note.—For simplicity, absolute values × 100 are provided and values less than 60 are omitted. Correlations between ordination methods are not given. J, Jaccard; SD, Sørensen-Dice; RR, Russell and Rao; SM, Simple Matching; RT, Rogers-Tanimoto; O, Ochiai; P, Phi; Y, Yule.

1; table 3). Jaccard and Sørensen-Dice coefficients produce nearly identical dendrograms. This pattern is also evident from dendrograms derived from Simple Matching and Rogers-Tanimoto coefficients. These two pairs of coefficients differ in the relative weighting of *a*, *b*, *c*, or *d* values, which appears to have limited impact on subsequent analyses.

The dependence of cluster analysis on size was also evident from the chaining of species in the dendrograms (fig. 1). The nearly identical dendrograms from the Jaccard and Sørensen-Dice coefficients produce species clusters mirroring the size gradient (see also table 3). This effect was quite apparent in the Russell and Rao dendrogram, where rare species chained into the existing core group. Simple Matching and Rogers-Tanimoto coefficients clustered species from "seed" groups formed around rare or ubiquitous species. This was due to the inclusion of joint absences in the coefficients' numerator; these dendrograms are thus unique (table 3). The dendrogram based on Ochiai's coefficient also appears to summarize a size effect (i.e., the cophenetic correlation with the Jaccard dendrogram is 0.971 and the consensus index is 0.826; table 3), whereas the Phi and Yule's coefficients produce different dendrograms.

This size dependence in cluster analysis is also evident in published studies, but the size artifact has not been previously identified (see, e.g., Harvey 1978, 1981; Somers and Harvey 1984; Nemec and Brinkhurst 1987). Consequently, many previous interpretations of "species associations" may actually describe a methodological artifact (Strauss 1982). We believe that size dependence per se simply reflects whether or not the similarity coefficient implicitly incorporates a centering transformation. Noy-Meir (1971) identified a similar phenomenon in principal-components analysis, where the analysis of non-centered data (i.e., nodal analysis) produced a first principal component ordering observations along an axis joining the origin and the centroid of the data (see also Noy-Meir 1973*a*). This method is also recognized as non-centered principal-components analysis (Hinch and Somers 1987), and similar effects emerge in PCoA and NMDS when non-centered similarity coefficients are used (for discussions of centering, see Dagnelie 1965; Orlóci 1967).

Ochiai's coefficient incorporates a centering translation described as the chord distance by Orlóci (1967). This differs from a centered coefficient (i.e., an association coefficient) in ranging from 0 to 1.0. Additionally, although the first PCoA axis was not correlated with frequency of occurrence, the third axis was (table 4). Consequently, Ochiai's coefficient may simply shift the size effect from the first axis to some subsequent axis. Interestingly, no size dependence is apparent in the NMDS solution.

The two association coefficients (i.e., Phi and Yule), which center data with the $ad - bc$ numerator, show no size dependence in the first two PCoA and NMDS axes. However, the third axis in the PCoA of the Phi coefficient is correlated with frequency of occurrence (table 4), just like the third axis associated with Ochiai's coefficient (see also table 6). These two association coefficients and Ochiai's coefficient incorporate centering transformations that reduce the importance of the size effects (i.e., shift size to subsequent axes). Apparently, the greatest effect of choosing a non-centered similarity coefficient is manifested by cluster-analysis procedures in which the dendrogram orders individuals along a single size axis. Size effects in ordinations are less troublesome, but failure to recognize size axes may jeopardize interpretations.

Obviously, generalized conclusions regarding the implications of selecting similarity coefficients on the basis of the results of this study alone would be premature. But it appears that the major difference between co-occurrence and association coefficients is the reduced emphasis of size dependence whenever the coefficient incorporates a centering transformation. In this study, multivariate summaries of similarity matrices faithfully reproduce a gradient correlated with the frequency of occurrence when non-centered similarities are used. This gradient appears in cluster analysis, PCoA, and NMDS. We fear that this size effect has been previously unnoticed; yet size may constitute a major portion of current interpretations.

SUMMARY

Data on the presence or absence of 25 fish species in a survey of 52 lakes from the watersheds of the Black and Hollow rivers of south-central Ontario were analyzed with eight similarity coefficients. Comparisons were made of Jaccard, Ochiai, Phi, Rogers-Tanimoto, Russell and Rao, Simple Matching, Sørensen-Dice, and Yule similarity coefficients using results from R-mode cluster analysis, principal-coordinates analysis (PCoA), and nonmetric multidimensional scaling. Coefficients were grouped into those representing measures of co-occurrence and those measuring association. Coefficients of co-occurrence (i.e., Jaccard, Rogers-Tanimoto, Russell and Rao, Simple Matching, and Sørensen-Dice) incorporate information associated with the frequency of occurrence of the fish species analyzed. Dendrograms faithfully revealed this size effect. Similarly, first axes of PCoA were linear or curvilinear functions of species' frequency of occurrence. Measures of association (i.e., Phi and Yule) and Ochiai's coefficient were less affected by the frequency of occurrence. The first axes of PCoA, based on centered coefficients (i.e., Phi, Yule, and Ochiai), were highly correlated with the second axes from ordinations using co-occurrence coefficients. The second axes from analyses of centered coefficients were correlated with the third axes based on non-centered measures.

We propose that co-occurrence coefficients reflect a general size effect similar to that commonly found in principal-components analysis. Measures of association and Ochiai's coefficient incorporate implicit centering transformations that reduce the size influence associated with the frequency of occurrence. Cluster analyses using co-occurrence coefficients are most susceptible to this size effect. We believe that the interpretations of many dendrograms fail to recognize size effects that arise from employing non-centered similarity coefficients (e.g., Strauss 1982; Nemec and Brinkhurst 1987). Additionally, arguments contrasting phenetic and phylogenetic methods may unknowingly debate the utility of centered versus non-centered coefficients, since the size effect undoubtedly contributes to the apparent strength of phylogenetic approaches.

ACKNOWLEDGMENTS

We thank A. Baker, R. Hansell, L. Orlóci, J. Podani, and three anonymous reviewers for comments on the manuscript. Data were collected with the help of I. Creed, A. Harvey, K. Keenleyside, J. Rosenfeld, and N. Salisbury. Portions of this research were funded by grants from the Ontario Ministry of Natural Resources Renewable Resources Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada to H.H.H., a University of Toronto Fellowship to D.A.J., and an NSERC postdoctoral fellowship to K.M.S.

LITERATURE CITED

Austin, M. P. 1976. Performances of four ordination techniques assuming three different nonlinear response models. Vegetatio 42:11–21.

Baroni-Urbani, C., and M. W. Buser. 1976. Similarity of binary data. Syst. Zool. 25:251–259.

Brown, S. D. 1969. Grouping plankton samples by numerical analysis. Hydrobiologia 33:289–301.

Cailliez, F., and J.-P. Pagès. 1976. Introduction à l'analyse des données. Société de Mathématiques Appliquées et de Sciences Humaines, Paris.

Cheetham, A. H., and J. E. Hazel. 1969. Binary (presence-absence) similarity coefficients. J. Paleontol. 43:1130–1136.

Clifford, H. T., and W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York.

Clymo, R. S. 1980. Preliminary survey of the peat-bog Hummell Knowe moss using various numerical methods. Vegetatio 42:129–148.

Cole, L. C. 1949. The measurement of interspecific association. Ecology 30:411–424.

Czekanowski, J. 1913. Zarys metod statystycznycg w zastosowaniu do antropologii. Travaux de la Société des Sciences de Varsovie III. Classes des sciences mathématiques et naturelles, no. 5.

Dagnelie, P. 1965. L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques: principes fondamentaux. Biometrics 21:345–361.

Del Moral, R. 1980. On selecting indirect ordination methods. Vegetatio 42:75–84.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297–302.

Everitt, B. S. 1979. Unresolved problems in cluster analysis. Biometrics 35:169–181.

Farris, J. S. 1979. On the naturalness of phylogenetic classification. Syst. Zool. 28:200–214.

Fasham, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology 58:551–561.

Fienberg, S. E., and J. P. Gilbert. 1970. The geometry of a two by two contingency table. J. Am. Stat. Assoc. 65:694–701.

Gauch, H. G., Jr. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Gauch, H. G., Jr., R. H. Whittaker, and T. R. Wentworth. 1977. A comparative study of reciprocal averaging and other ordination techniques. J. Ecol. 65:157–174.

Gordon, A. D. 1987. A review of hierarchical classification. J. R. Stat. Soc. A 150:119–137.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

———. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Gower, J. C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3:5–48.

Hart, G. 1983. The occurrence of multiple UPGMA phenograms. Pages 254–258 in J. Felsenstein, ed. Numerical taxonomy. NATO ASI Ser. G, Ecol. Sci. 1. Springer-Verlag, Berlin.

Harvey, H. H. 1978. The fish communities of the Manitoulin Island lakes. Int. Ver. Theor. Angew. Limnol. Verh. 20:2031–2038.

———. 1981. Fish communities of the lakes of the Bruce Peninsula. Int. Ver. Theor. Angew. Limnol. Verh. 21:1222–1230.

Hinch, S. G., and K. M. Somers. 1987. An experimental evaluation of the effect of data centering, data standardization, and outlying observations on principal components analysis. Coenoses 2:19–23.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biol. Rev. Camb. Philos. Soc. 57:669–689.

Hughes, C. P. 1973. Analysis of past faunal distributions. Pages 221–230 in D. H. Tarling and S. K. Runcorn, eds. Implications of continental drift to the earth sciences. Academic Press, New York.

Hurlbert, S. H. 1969. A coefficient of interspecific association. Ecology 50:1–9.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. Soc. Vaudoise Sci. Nat. 37:547–579.

Jackson, D. A. 1988. Fish communities in lakes of the Black and Hollow River watersheds, Ontario. Master's thesis, University of Toronto, Toronto.

Jain, N. C., A. Indrayan, and L. R. Goel. 1986. Monte Carlo comparisons of six hierarchical clustering methods on random data. Pattern Recogn. 19:95–99.

Janowitz, M. F. 1979. A note on phenetic and phylogenetic classification. Syst. Zool. 28:197–199.

———. 1980. Similarity measures on binary data. Syst. Zool. 29:342–359.

Jolicoeur, P., and J. E. Mosimann. 1960. Size and shape variation in the painted turtle: a principal component analysis. Growth 24:339–354.

Kenkel, N. C., and T. Booth. 1987. A comparison of presence-absence resemblance coefficients for use in biogeographical studies. Coenoses 2:25–30.

Kenkel, N. C., and L. Orlóci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67:919–928.

Kruskal, J. B. 1964*a*. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27.

———. 1964*b*. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115–129.

Lamont, B. B., and K. J. Grant. 1979. A comparison of twenty-one measures of site dissimilarity. Pages 101–126 in L. Orlóci, C. R. Rao, and W. M. Stiteler, eds. Multivariate methods in ecological work. International Co-operative Publishing House, Fairland, Md.

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

Legendre, P., and V. Legendre. 1984. The postglacial dispersal of freshwater fishes in the Quebec Peninsula. Can. J. Fish. Aquat. Sci. 41:1781–1802.

Michael, E. L. 1920. Marine ecology and the coefficient of association: a plea in behalf of quantitative ecology. J. Ecol. 8:54–59.

Nemec, A. F. L., and R. O. Brinkhurst. 1987. A comparison of methodological approaches to the subfamilial classification of the Naididae (Oligochaeta). Can. J. Zool. 65:691–707.

Noy-Meir, I. 1971. Multivariate analysis of the semi-arid vegetation in south-eastern Australia: nodal ordination by component analysis. Proc. Ecol. Soc. Aust. 6:159–193.

———. 1973*a*. Data transformations in ecological ordination. I. Some advantages of non-centering. J. Ecol. 61:329–341.

———. 1973*b*. Divisive polythetic classification of vegetation data by optimized division on ordination components. J. Ecol. 61:753–760.

Ochiai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jpn. Soc. Sci. Fish. (Nihon Suisan Gakkaishi) 22:526–530.

Orlóci, L. 1967. Data centering: a review and evaluation with reference to component analysis. Syst. Zool. 16:208–212.

———. 1978. Multivariate analysis in vegetation research. 2d ed. Junk, The Hague.

Peters, J. A. 1971. A new approach in the analysis of biogeographic data. Smithson. Contrib. Zool. 107:1–28.

Pielou, E. C. 1984. The interpretation of ecological data: a primer on classification and ordination. Wiley, New York.

Rice, J., and R. J. Belland. 1982. A simulation study of moss floras using Jaccard's coefficient of similarity. J. Biogeogr. 9:411–419.

Rogers, D. J., and T. T. Tanimoto. 1960. A computer program for classifying plants. Science (Wash., D.C.) 132:1115–1118.

Rohlf, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. 59:131–144.

Rohlf, F. J., J. Kishpaugh, and D. Kirk. 1982. NT-SYS, numerical taxonomy system of multivariate statistical programs. State University of New York, Stony Brook.

Russell, P. F., and T. R. Rao. 1940. On habitat and association of species of anopheline larvae in south-eastern Madras. J. Malaria Inst. India 3:153–178.

Sepkowski, J. J., and M. A. Rex. 1974. Distribution of freshwater mussels: coastal rivers as biogeographic islands. Syst. Zool. 23:165–188.

Shepard, R. N. 1974. Representation of structure in similarity data: problems and prospects. Psychometrika 39:373–421.

Sibson, R. 1979. Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. J. R. Stat. Soc. B, Methodol. 41:217–229.

Simberloff, D. S., and E. F. Connor. 1979. Q-mode and R-mode analyses of biogeographic distributions: null hypotheses based on random colonization. Pages 123–138 *in* G. P. Patil and M. Rosenzweig, eds. Contemporary quantitative ecology and related econometrics. International Co-operative Publishing House, Fairland, Md.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. Freeman, San Francisco.

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38:1409–1438.

Somers, K. M. 1986. Multivariate allometry and removal of size with principal components analysis. Syst. Zool. 35:359–368.

Somers, K. M., and H. H. Harvey. 1984. Alteration of fish communities in lakes stressed by acid deposition and heavy metals near Wawa, Ontario. Can. J. Fish. Aquat. Sci. 41:20–29.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. K. Dan. Vidensk. Selsk. Biol. Skr. (Copenhagen) 5:1–34.

Stephenson, W., and W. T. Williams. 1971. A study of the benthos of soft bottoms, Sek Harbour, New Guinea, using numerical analysis. Aust. J. Mar. Freshwater Res. 22:11–34.

Strauss, R. E. 1982. Statistical significance of species clusters in association analysis. Ecology 63:634–639.

Williams, W. T., H. T. Clifford, and G. N. Lance. 1971. Group-size dependencies: a rationale for choice between numerical classifications. Comput. J. 14:157–162.

Yule, G. U. 1900. On the association of attributes in statistics. Philos. Trans. R. Soc. A 194:257–319.

———. 1912. On the methods of measuring association between two attributes. J. R. Stat. Soc. 75:579–642.