COENOSES 2(1): 25-30, 1987
© 1987 by Edizioni LINT Trieste

A COMPARISON OF PRESENCE-ABSENCE RESEMBLANCE COEFFICIENTS FOR USE IN BIOGEOGRAPHICAL STUDIES

N.C. Kenkel and T. Booth, Department of Botany, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

Keywords: Biogeography, Individual, Differences, Nonmetric, Multidimensional, Scaling, Ordination, Presence-absence, Resemblance

Abstract: A comparison of thirteen presence-absence resemblance coefficients was undertaken based on nonmetric multidimensional scaling of a data set recording the distribution of marine fungal species over 31 sites worldwide. While all coefficients successfully recovered a major gradient in mean seawater temperature, some differences in the ordination scattergrams were nonetheless apparent. The weights space of individual differences scaling (INDSCAL) indicated that coefficient interrelationships reflect (a) whether mutual absence is used in calculating resemblance, (b) whether multiplicative terms are incorporated, and (c) the degree of intercoefficient monotonicity. An empirical assessment of ordination interpretability, together with considerations of the relationship between resemblance and ecological distance as indicated by INDSCAL Shepard diagrams, suggested the utility of the Baroni-Urbani coefficient in biogeographical studies. The Ochiai coefficient, along with the Jaccard and its variants, also produced interpretable results.

Introduction

While a number of presence-absence resemblance coefficients have been proposed for determining pairwise relationships among sites (Cheetham and Hazel 1969; Anderberg 1973; Goodall 1978), few studies have assessed their relative merits and limitations. This stems largely from the fact that many of the suggested coefficients lack an underlying statistical or mathematical basis, precluding a rigorous, formalized approach for choosing among them (Cheetham and Hazel 1969). Coefficients must therefore be assessed through empirical studies in which the fulfillment of prior expectations plays an important role (Huhta 1979).

Because the various resemblance coefficients emphasize different aspects of the underlying data structure, coefficient choice may have a strong influence on data analysis and subsequent interpretation (Baroni-Urbani and Buser 1976). Indeed, the measurement of interspecific relationships is a basic step in data analysis, since the multivariate methods such as ordination and cluster analysis are generally based on the analysis of a resemblance matrix rather than the raw data itself.

This work undertakes an empirical comparison of 13 presence-absence coefficients, based on nonmetric multidimensional scaling ordinations of a single data set. Comparisons are restricted to non-centred, non-probabilistic coefficients (Dagnelie 1960; Goodall 1978) which have been utilized or suggested as being of potential utility in biogeographical studies. Emphasis is placed on determining interrelationships among the coefficients, and on examining the behaviour of the coefficients with regard to their success in recovering underlying data structure.

Methods

Data set

The data set records the presence-absence of 68 lignicolous marine fungal species in 31 sites worldwide (Fig. 1). Data were collected from the literature (see Booth and Kenkel 1986), though only those species occurring in at least five sites were included. The data set is typical of presence-absence biogeographical data, with approximately two-thirds zero entries.

Previous analyses based on classification and ordination utilizing the Ochiai (1957) coefficient indicated the overall importance of mean seawater temperature in dictating the worldwide distribution of lignicolous marine fungal species, with three major site types revealed by cluster analysis: tropical, subtropical-temperate, and temperate (Booth and Kenkel 1986). Within these types, differences attributable to temperature and salinity variability were often apparent.

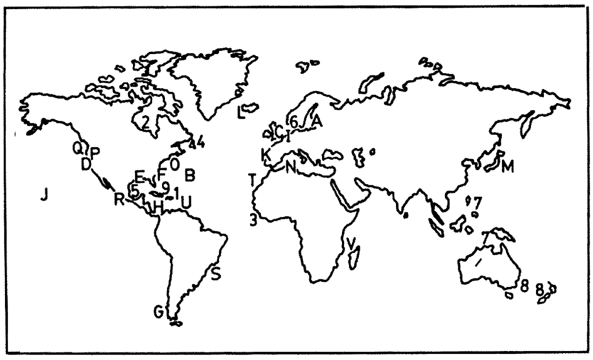

Fig. 1. Locations of the 31 sample sites (coded 1-9, A-V).

$$C1 = \frac{a}{a + b + c + d} \quad \text{(Russell/Rao)}$$

$$C2 = \frac{a + d}{a + 2b + 2c + d} \quad \text{(Rogers/Tanimoto)}$$

$$C3 = \frac{a + d}{a + b + c + d} \quad \text{(Sokal/Mitchener)}$$

$$C4 = \frac{2a + 2d}{2a + b + c + 2d} \quad \text{(Sokal/Sneath 1)}$$

$$C5 = \left[\left(\frac{a}{a + b}\right)\left(\frac{a}{a + c}\right)\right]^{1/2} \quad \text{(Ochiai)}$$

$$C6 = \frac{1}{2}\left(\frac{a}{a + b} + \frac{a}{a + c}\right) \quad \text{(Kulcznski)}$$

$$C7 = \frac{a}{a + b + c} \quad \text{(Jaccard)}$$

$$C8 = \frac{2a}{2a + b + c} \quad \text{(Sorensen)}$$

$$C9 = \frac{a}{a + 2b + 2c} \quad \text{(Sokal/Sneath 2)}$$

$$C10 = \left[\left(\frac{a}{a + b}\right)\left(\frac{a}{a + c}\right)\left(\frac{d}{b + d}\right)\left(\frac{d}{c + d}\right)\right]^{1/2} \quad \text{(Anderberg 1)}$$

$$C11 = \frac{[ad]^{1/2} + a}{[ad]^{1/2} + a + b + c} \quad \text{(Baroni-Urbani)}$$

$$C12 = \frac{1}{4}\left(\frac{a}{a + b} + \frac{a}{a + c} + \frac{d}{b + d} + \frac{d}{c + d}\right) \quad \text{(Anderberg 2)}$$

$$C13 = \frac{2a}{ab + ac + 2bc} \quad \text{(Mountford)}$$

Table 1. Names and formulae for the thirteen coefficients compared. Symbols C1 - C13 are used in the main text.

Coefficients compared

When presence-absence of species within sites is recorded, the relationship between any two sites can be expressed in the form of a 2×2 contingency table:

		Site 2	
		+	−
Site 1	+	*a*	*b*
	−	*c*	*d*

where $a =$ number of species common to both sites, $b =$ number of species present only at site 1, $c =$ number of species present only at site 2, and $d =$ number of species absent from both sites. Presence-absence coefficients are normally expressed as a function of the elements *a*, *b*, *c*, and *d* (Goodall 1978).

The thirteen resemblance coefficients compared in this study are given in Table 1. Original references, which were used in naming the coefficients, can be found in most review articles (*e.g.* Cheetham and Hazel 1969; Goodall 1978). A general discussion of each follows:

C1 - Russell/Rao. The probability that two sites score positively (mutual presence) on a randomly selected variable is measured by this coefficient. The lack of the *d* term in the numerator, and its presence in the denominator, makes it sensitive to large values of *d*.

C2 - Rogers/Tanimoto. This coefficient includes the *d* term in both the numerator and denominator, and gives unmatched pairs double weight. Anderberg (1973) points out that it represents the ratio of the number of classes held in common by two sites to the total number of classes represented.

C3 - Sokal/Mitchener. Also known as the 'Simple Match', this coefficient represents the probability that two sites have the same score (mutual presence or absence) on a randomly chosen variable.

C4 - Sokal/Sneath 1. Though similar to C3, this coefficient gives double weight to matches (mutual presences and absences). It is monotonic with coefficients C2 and C3.

C5 - Ochiai. This coefficient is a scalar product measuring the angle between two vectors (Orlóci 1978). Anderberg (1973) shows that it represents the geometric mean of the conditional probabilities associated with the element *a*.

C6 - Kulczynski. The arithmetic mean of conditional probabilities associated with the element *a*, rather than the geometric mean (cf. C5), defines this coefficient.

C7 - Jaccard. The conditional probability that both sites score positively (mutual presence) on a randomly chosen variable, once mutual absences are discarded, is measured by this coefficient.

C8 - Sorensen. Because it gives double weight to mutual presences, this coefficient has no probabilistic interpretation (Anderberg 1973). It is monotonic with coefficients C7 and C9. Hall (1969) suggests that double weighting of the element *a* is justified by the fact that in mismatches (*b* and *c*) absence is just as uninformative as in mutual absences.

C9 - Sokal/Sneath 2. This coefficient is a variant of C8 which gives double weight to mismatches rather than mutual presences; again, it has no probabilistic interpretation.

C10 - Anderberg 1. The geometric mean of conditional probabilities associated with the elements *a* and *d* is measured by this coefficient. It is therefore a variant of the Ochiai coefficient (C5) which includes the *d* term.

C11 - Baroni-Urbani. This empirically derived coefficient was developed to incorporate the *d* term without giving it excessive weight (Baroni-Urbani and Buser 1976). It lacks a probabilistic interpretation.

C12 - Anderberg 2. The arithmetic mean of conditional probabilities associated with the elements *a* and *d*, rather than the geometric mean (cf. C10), defines this coefficient. It is therefore a version of the Kulczynski coefficient (C6) which incorporates the *d* term.

C13 - Mountford. The formula given (Table 1) is an approximation of a coefficient suggested by Mountford (1962) which was developed to render coefficient values independent of sample size under the assumption that the distribution of the underlying population is logarithmic (see Orlóci 1978).

In all cases distance versions of the similarity measures were used in obtaining site ordinations. Most distances were obtained by computing one-complements of the similarity values. Exceptions were the Ochiai coefficient (C5), for which distances were obtained via the transformation

$$D5 = [2(1 - C5)]^{1/2}$$

giving a chord distance (Orlóci 1967) which ranges between 0 and 1.414. The Mountford index (C13), which has no fixed upper bound, was transformed by the exponential function suggested by Orlóci (1978) to obtain a distance measure ranging between 0 and 1.

Nonmetric multidimensional scaling

This ordination method seeks an arrangement of individuals in a reduced metric ordination space, such that the dissimilarities **d** calculated in this space are as closely monotonic as possible to the original distances $\boldsymbol{\delta}$ calculated in variable space (Kruskal 1964 a, b). Since monotonicity is the only constraint, the method can accept as input virtually any resemblance matrix (Wish and Carroll 1982). In this study 13 separate ordinations, one for each resemblance matrix, were obtained using the program ALSCAL (Takane and Young 1977). In each case a two-dimensional solution was specified, and the input starting configuration was based on metric multidimensional scaling.

Individual differences scaling

This method (INDSCAL) is a generalization of two-way multidimensional scaling which permits the simultaneous analysis of a series of resemblance matrices (Carroll and Chang 1970). Also called weighted multidimensional scaling (Schiffman *et al.* 1981), the method differs from ordinary multidimensional scaling by substituting a weighted Euclidean metric for the normal unweighted form. This permits examination of variation within 'subjects' (resemblance matrices) in the context of a shared ordination space. Thus two spaces are defined, the so-called 'stimulus space' which is essentially an averaged configuration based on the entire set of resemblance matrices, and a 'weights space' which summarizes interrelationships among the resemblance matrices. Each matrix is represented by a weight vector which indicates its importance on each dimension of the stimulus space.

INDSCAL assumes that the individual spaces (defined by the resemblance matrices) are related to one another by linear transformations of the stimulus space. Given that the resemblance matrices in this study are based on a common data set, and that different resemblance structures can be expected to be approximately monotonic (Clifford and Stephenson 1975), such an assumption is probably tenable. An important consequence of the assumption is that, unlike nonmetric multidimensional scaling, individual differences scaling axes are uniquely determined (Wish and Carroll 1982).

The INDSCAL option of program ALSCAL, at the ordinal (non-metric) measurement level, was used to obtain both a common stimulus space and a weights space for assessing interrelationships among the 13 coefficients.

Shepard diagrams

Shepard (1962) suggested plotting dissimilarities (**d**) against distances ($\boldsymbol{\delta}$) to determine the 'goodness-of-fit' of multidimensional scaling results. Alternatively, one could plot disparities ($\hat{\mathbf{d}}$) against distances. Disparities represent modified **d** values chosen to be as nearly equal to them as possible under the constraint that the plot of $\hat{\mathbf{d}}$ against $\boldsymbol{\delta}$ be monotonic (Kruskal 1964 a). In general, the smoother the curve, the fewer the number of ties and the better the fit. Gauch *et al.* (1981) have suggested that the Shepard diagram may be interpreted ecologically as the relationship between resemblance structure and gradient distance. In such cases a linear relationship is to be preferred (Gauch 1973). In this study the 13 Shepard diagrams resulting from individual differences scaling were used to assess coefficient utility.

Results

Nonmetric multidimensional scaling

Scattergrams of the 31 sites resulting from each of the thirteen coefficients are presented in Figure 2. The stress value measures the overall 'badness-of-fit' of the ordination configuration relative to the input distance matrix (Kruskal 1964 a, b). The results indicate that coefficients C2 - C4 give marginally better fits, though differences are slight.

For ease of interpretation, each scattergram has been partitioned into three major site types, delineated using sum of squares agglomerative cluster analysis based on the Ochiai coefficient (see Booth and Kenkel 1986). The types are: I = temperate; II = subtropical-temperate; and III = tropical. While all the ordinations separate out the three types well, for some coefficients (particularly C2 - C4) discrimination appears to be weaker. Within the three site types differences in some of the ordination results are apparent.

Strong similarities among certain ordination results are also revealed. Coefficients C5 - C8 produce nearly identical ordinations, with C9 also quite similar. Ordination results from C2 - C4 are indistinguishable, while those produced by using coefficients C10 - C13 are also quite similar. The results from C1 are most similar to those obtained using C5 - C9.

Individual differences scaling

The stimulus space derived from simultaneous analysis of the 13 coefficients is presented in Figure 3. The first axis reflects a gradient in mean seawater temperature, from the cool temperate sites at the left to the tropical sites at the right. The second axis reflects variation within each of the three major groups. The overall importance of the axes (the mean of squared 'subject' weights) are 0.722 and 0.065 respectively, indicating the overriding importance of the first.

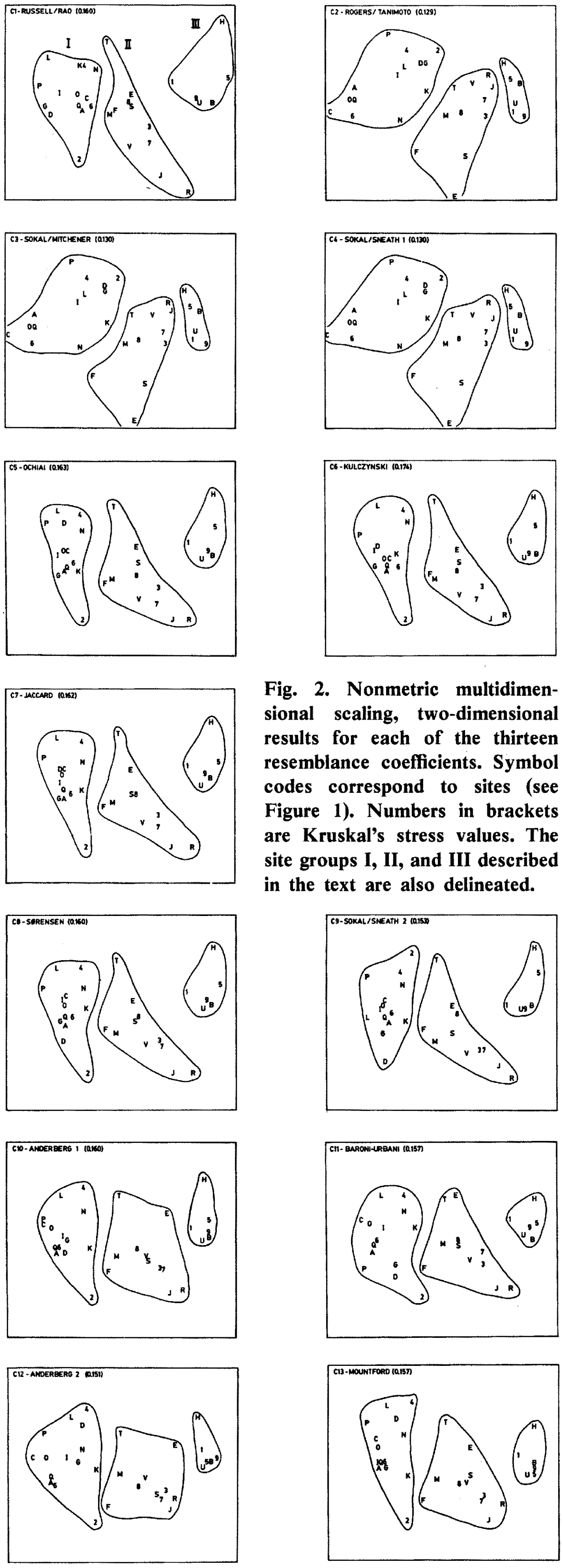

Fig. 2. Nonmetric multidimensional scaling, two-dimensional results for each of the thirteen resemblance coefficients. Symbol codes correspond to sites (see Figure 1). Numbers in brackets are Kruskal's stress values. The site groups I, II, and III described in the text are also delineated.

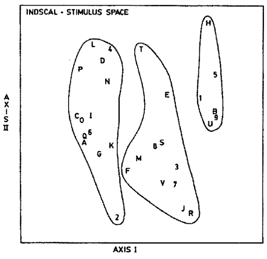

Fig. 3. Individual differences scaling, stimulus space for the 31 sites.

The weights for each of the 13 resemblance coefficients are presented graphically in Figure 4. Coefficients C1 - C4 have the lowest weights on the first axis, while the highest weights are shown by coefficients C10 - C13. The results suggest the recognition of five groupings of the 13 coefficients:

1. Coefficient C1 occurs by itself, and has the lowest weight on axis I. This coefficient is unique in that it contains the *d* term in the denominator but not the numerator.
2. Virtually identical weights result from coefficients C2 - C4. These are simple additive coefficients which contain the *d* term in both the numerator and denominator.
3. The five coefficients C5 - C9 are also closely related. They share the common property of excluding the *d* term from both the numerator and denominator.
4. Coefficients C10, C11, and C13 have similar weightings on the two axes. They appear to have little in common other than the presence of multiplicative terms in their calculation.
5. The weightings for the final coefficient (C12) indicate closest affinity with C10, C11, and C13.

Shepard diagrams

Figure 5 presents the INDSCAL Shepard diagrams (disparities vs. distances) for each of the 13 coefficients. Those showing a concave upward curve include C1, C9, and C13, and to a lesser extent C2, C5, C7, and C10. A

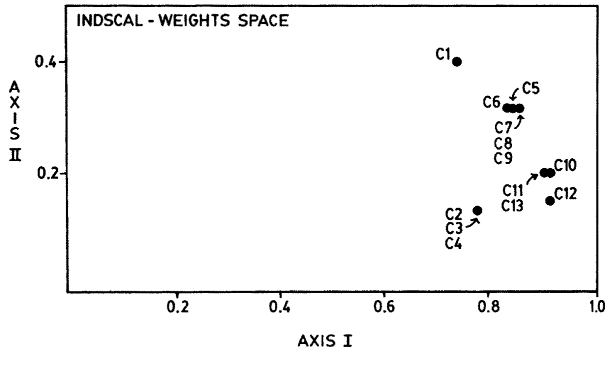

Fig. 4. Individual differences scaling, weights space for the thirteen coefficients.

concave downward curve is produced by C4. The others (C3, C6, C8, C11, and C12) are more or less linear in form.

Discussion

An overall consistency of the thirteen scattergrams is apparent, with all coefficients leading to the successful recovery of the major gradient in mean seawater temperature. Nonetheless, some differences are attributable to the coefficients differentially emphasizing specific aspects of the underlying data structure. To examine this further, a subjective assessment of the scattergrams was undertaken using biological, hydrological, and geographical integrity, group recognition, and pattern interpretability as criteria. For each of these five criteria, scattergrams were ranked as poor, fair, or good based on prior knowledge of the sites (see Booth and Kenkel 1986 and references therein). The results suggested that coefficient C11, and to a lesser extent C5, produced the most satisfactory scattergrams. Moderately interpretable results were produced by coefficients C6 - C9. The remaining coefficients (C1 - C4, C10, C12, and C13) gave results which were weak in one or more of the five assessment criteria.

Interrelationships among the coefficients compared have also been clarified. Coefficients C6 - C9, which all exclude the *d* term in assessing similarity, produced virtually identical ordination results and had similar INDSCAL weights. Of these, C8 showed the strongest linear relationship between disparities and distances. The INDSCAL weights for coefficient C5 were similar to those of C6 - C9, though its scattergram was somewhat more interpretable. Coefficient C1, which includes the *d* term in the denominator but not the numerator, had the lowest weight on the first INDSCAL axis and produced a Shepard diagram indicating a poor fit and a nonlinear trend. Coefficients C2 - C4 produced very similar scattergrams, which is to be expected given that they are monotonic (Anderberg 1973; Baroni-Urbani and Buser 1976). The Shepard diagrams for these coefficients indicated relatively poor fits. It is interesting to note that coefficients C10 - C13 produced similar INDSCAL weights, though on first sight they appear to have little in common (Table 1). The ordination scattergrams produced by these four coefficients were somewhat different, however. Coefficients C10 and particularly C13 produced nonlinear Shepard diagrams, while those for C11 and C12 were linear. Wolda (1981) noted that coefficient C13 takes on low values for most of its range before increasing dramatically, which would account for the highly nonlinear relationship between disparities and distances found in this study.

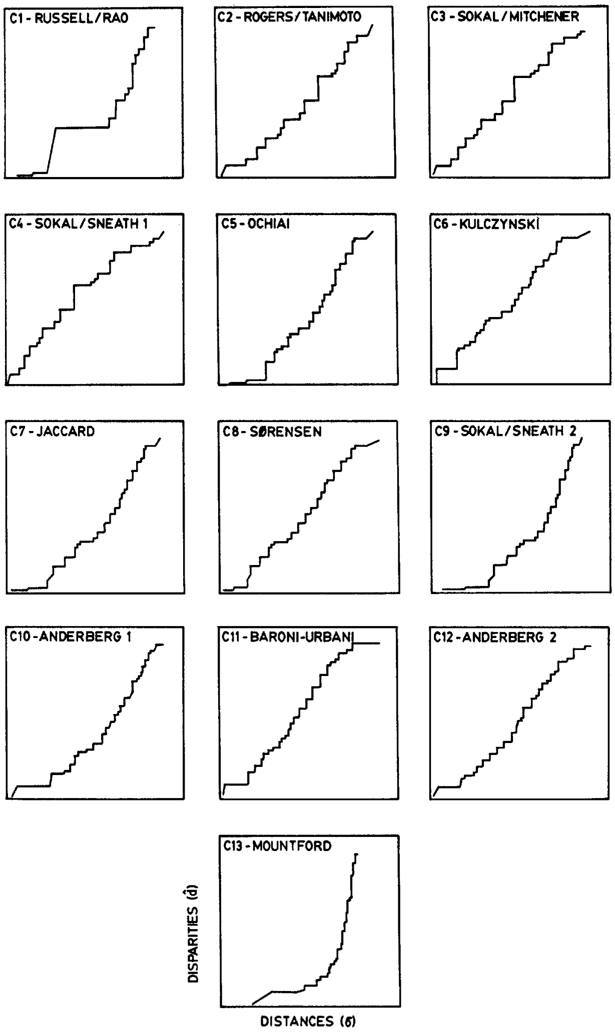

Fig. 5. Individual differences scaling, Shepard diagrams (disparities vs. distances) for the thirteen coefficients.

In assessing the utility of presence-absence coefficients, the question of whether to include the *d* term in the numerator has received much attention. If the data is two-state nominal the *d* term clearly must be included (Sneath and Sokal 1973; Baroni-Urbani and Buser 1976). For presence-absence data the issue is less clear. Species absence may reflect either stochastic phenomena or the fact that the niche is occupied by a competing species (Green 1971; Legendre and Legendre 1983). Pielou (1984) makes similar comments, pointing out that while species presence offers definite information regarding its ability to persist at a site, absence may be attributable either to probabilistic or deterministic factors. Campbell (1978) concluded that mutual absence of species should not be used in assessing stand interrelationships, since the absence of two species may be due to completely different ecological constraints (see also Cairns 1974; Green 1979). Furthermore, in the case of rare species mutual absence is largely a stochastic phenomenon which has little ecological relevance (Legendre and Legendre 1983). Thus for presence-absence data sets containing many zeroes, a high similarity among individuals attributable to the mutual absence of the majority of species may arise if the *d* term is used in the calculation of resemblance (Baroni-Urbani and Buser 1976; Green 1979). Goodall (1978) takes a similar stand, but points out that the mutual presence of ubiquitous species may be equally uninformative (see also Greig-Smith 1983). Further complicating the issue is the contention (Clifford and Stephenson 1975) that exclusion of the *d* term may lead to the loss of potentially important information regarding the joint absence of ubiquitous species. Baroni-Urbani and Buser (1976) feel that criticism lodged against inclusion of the *d* term stems from the fact that most of the coefficients suggested have «... improperly used this type of information». They state that resemblance must take into account the *d* term, but show that coefficient C3 and related forms are inappropriate since excessive weight is given to mutual absences. They suggest instead using the product *ad*, but recommend its square root to ensure that the value is of the same order of magnitude as the other elements. Our results indicate that their coefficient (C11) may be useful in exploratory biogeographical studies; indeed, it appears to offer a compromise between simple coefficients incorporating the *d* term and those which exclude mutual absences entirely. Further empirical studies are clearly required, however, and considerations should also be given to the distributional properties of the coefficients as well as the effect of sample size (Goodall 1978; Wolda 1981) on coefficient values.

Acknowledgements. Financial support from the University of Manitoba Research Committee to N.C.K., and the Natural Sciences and Engineering Research Council (grant A0303) to T.B., is gratefully acknowledged.

REFERENCES

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

BARONI-URBANI, C. and M.W.BUSER. 1976. Similarity of binary data. Syst. Zool. 25: 251-259.

BOOTH, T. and N. KENKEL, 1986. Ecological studies of lignicolous marine fungi: a distribution model based on ordination and classification. Pages 297-310. *In*: S.T. Moss (ed.), *The Biology of Marine Fungi.* Cambridge University Press, Cambridge.

CAIRNS, J. 1974. Indicator species vs. the concept of community structure as an index of pollution. Water Resour. Bull. 10: 338-347.

CAMPBELL, B.M. 1978. Similarity coefficients for classifying releves. Vegetatio 37: 101-109.

CARROLL, J.D. and J.J. CHANG. 1970. Analysis of individual differences in multidimensional scaling via a N-way generalization of Eckart-Young decomposition. Psychometrika 35: 283-319.

CHEETHAM, A.H. and J.E. HAZEL. 1969. Binary (presence-absence) similarity coefficients. J. Palentol. 43: 1130-1136.

CLIFFORD, H.T. and W. STEPHENSON. 1975. *An Introduction to Numerical Classification.* Academic Press, New York. 229 p.

DAGNELIE, P. 1960. Contribution à l'étude des communautés végétales par l'analyse factorielle. Bull. Serv. Carte Phytogéogr., Sér. B. 5: 7-71, 93-195.

GAUCH, H.G. 1973. The relationship between sample similarity and ecological distance. Ecology 54: 618-622.

GAUCH, H.G., R.H. WHITTAKER, and S.B. SINGER. 1981. A comparative study of nonmetric ordinations. J. Ecol. 69: 135-152.

GOODALL, D.W. 1978. Sample similarity and species correlation. Pages 99-149 *In*: Whittaker, R.H. (ed.), *Ordination of Plant Communities.* Junk, The Hague.

GREEN, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: Bivalve molluscs of central Canada. Ecology 52: 543-556.

GREEN, R.H. 1979. *Sampling Design and Statistical Methods for Environmental Biologists.* Wiley, New York. 257 p.

GREIG-SMITH, P. 1983. *Quantitative Plant Ecology.* 3rd ed. U. of California Press, Berkeley. 359 p.

HUHTA, V. 1979. Evaluation of different similarity indices as measures of succession in Arthropod communities of the forest floor after clear-cutting. Oecologia 41: 11-23.

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

KRUSKAL, J.B. 1964 b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29: 115-129.

LEGENDRE, L. and P. LEGENDRE. 1983. *Numerical Ecology.* Elsevier, Amsterdam. 419 p.

MOUNTFORD, M.D. 1962. An index of similarity and its application to classificatory problems. Pages 43-50 *In*: P.W. Murphy (ed.). *Progress in Soil Zoology.* Butterworth's, London.

ORLÓCI, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. 55: 193-206.

ORLÓCI, L. 1978. *Multivariate Analysis in Vegetation Research.* Junk, The Hague. 451 p.

PIELOU, E.C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination.* Wiley, New York. 263 p.

SCHIFFMAN, S.S., M.L. REYNOLDS, and F.W. YOUNG. 1981. *Introduction to Multidimensional Scaling. Theory, Methods, and Applications.* Academic Press, New York. 413 p.

SHEPARD, R.N. 1962. Analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27: 125-140, 219-246.

SNEATH, P.H.A. and R.R. SOKAL. 1973. *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* Freeman, San Francisco. 573 p.

TAKANE, Y. and F.W. YOUNG. 1977. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika 42: 7-67.

WISH, M. and J.D. CARROLL. 1982. Multidimensional scaling and its applications. Pages 317-345 *In*: P.R. Krishnaiah and L.N. Kanal (eds.). *Handbook of Statistics*, Volume II. North-Holland, Amsterdam.

WOLDA, H. 1981. Similarity indices, sample size and diversity. Oecologia 50: 296-302.

Manuscript received: March 1986