Vol. 5: 125–128, 1981 | MARINE ECOLOGY – PROGRESS SERIES Mar. Ecol. Prog. Ser. | Published May 31

# Similarity Indices in Community Studies: Potential Pitfalls

**Stephen A. Bloom**

Department of Zoology, University of Florida, Gainesville, Florida 32611, USA

ABSTRACT: Four common similarity indices used in multivariate descriptive techniques, such as classifications and trellis diagrams, are compared over a range of overlap from 100 to 10 % to a theoretical standard. Only the Bray-Curtis Index (also known as Czekanowski's Quantitative Index, Proportional Similarity and a variety of other names) was found to reflect accurately true similarity. The other indices (Canberra Metric, Morisita's and Horn's Information Theory) diverge greatly from one another and from the theoretical standard.

## INTRODUCTION

Multivariate techniques (and thus similarity indices) are standard analytical tools in community ecology. There is a wide variety of indices in use but no one is preferred (Boesch, 1977). The purpose of this paper is to point out that while a variety of indices vary between the same limits (0 to 1), they do not give comparable values for the same amount of actual similarity. This point has been made tacitly by others (Williams et al., 1973) but a quantitative and simultaneous comparison of the indices to one another and to a theoretical standard was not performed. Such a comparison reveals that there are major potential difficulties in cross-study comparisons, in the indiscriminant use of the term 'percent similarity' and in the interpretation of changes in similarity through time. The purpose here is not to present an exhaustive treatment of similarity indices but to sound a warning to users (especially of computer packages) who may not have delved into the voluminous literature dealing with similarity indices.

## METHODS

To compare the various indices, a common data base of known similarity is needed. The simplest data base available is a table of the area of a normal curve (Rohlf and Sokol, 1969, p. 158). If two normal curves overlap, the total area of overlap is twice the area from the intersection point of the curves through the tail of the distribution. Since a standard table of the area of a normal curve gives the area between the center and any given point, simple subtraction and multiplication will yield the area of overlap for any degree of separation between the overlapping curves.

A normal curve can be converted into a histogram by arbitrarily selecting some value (here set to 0.1 standard deviations) and dividing the curve into 60 such segments or resources ($-3.0$ to $+3.0$). The area within each block is calculated by determining the values at the edges of the block, e.g. $-3.0$ to $-2.9$, and using a table of the area of a normal curve to determine the area between those values. The area in each block can then be taken as the relative amount utilized for that resource, or as the 'counts' for that species at a station, after a frequency standardization.

Such histogram-normal curves can be calculated for any desired index. Operationally, this is equivalent to forming a 30-row $\times$ 90-column matrix in which the column of 60 positive area values successively moves to the right one column and down one row with all the other values in the matrix being set to zero, until the first positive value of the 30th column is in the same row as the 30th positive value of the first column. This is equivalent to a separation of the means of the normal curves represented by the 1st and 30th columns of 3.0 standard deviation units. Since the actual and precise area of overlap for each degree of separation can be



0171-8630/81/0005/0125/$ 02.00

determined simply from the table of the area of a normal curve, the actual similarity (or overlap) can be determined and compared to the index measures for the same degree of separation.

It is obvious that the choice of curves will control the absolute values of the results. Analyses of normal and log-normal curves were carried out and the qualitative results were identical. Since it is possible, using appropriate transformations, to convert one distribution to another, e.g. $y = \log(x)$, and (as will be pointed out later) the preferred index should perform equally well on any distribution due to its mathematical nature, only the results of the normal distribution are presented.

Four indices were selected for comparison. The first is known by a variety of names (Czekanowski's Index, Bray-Curtis Index, Schoener's Index, Least Common Percentage Index, Index of Affinity or Proportional Similarity; see Boesch, 1977 for a comprehensive review). Since the term 'Czekanowski's Index' is often used in marine studies (Field and McFarlane, 1968; Day et al., 1971; Field, 1971; Santos and Simon, 1974; Dauer and Simon, 1975; Santos and Bloom, 1980) and should be distinguished from the qualitative (presence/absence) form of the expression (sometimes called Dice's Coefficient), it will be referred to here as 'Czekanowski's Quantitative Index'. The other indices were Horn's Information Theory Index (Horn, 1966), Canberra Metric (Lance and Williams, 1967) and Morisita's Index as modified by Horn (Horn, 1966); see Boesch (1977) for a comprehensive review of these indices and a general literature review. Given the length and purpose of this note, an extensive review of the literature is not appropriate. Equations are presented in the Appendix. Given the conditions of the comparison, e.g. a frequency transformation such that all blocks total to 1.0 for each curve and each curve is identical, Morisita's modified index is mathematically identical to Levin's, Pianka's and MacArthur's measures of niche overlap (May, 1975). This is not to say that these measures are identical outside of the special case examined here.

A FORTRAN IV program, ORDANA (Bloom et al., 1977) was used to analyse the matrix for the 4 similarity indices. The calculated similarity values for the 4 indices between the first column of the matrix and the other 29 columns were plotted against the separation between the curves (Fig. 1A).

To facilitate index comparisons, the percent deviation from actual similarity was calculated by:

$$D_{ij} = 100.0\ (I_{ij} - T_{ij})/T_{ij}$$

where $D_{ij}$ = percent deviation; $I_{ij}$ = index value; $T_{ij}$ = actual value of overlap for the ith index and the jth separation of the curves (Fig. 1B).

## RESULTS

As is readily apparent in Fig. 1, the indices do not give comparable absolute values for a given degree of overlap. Each index generates a characteristic and distinct response to decreasing overlap. For moderate values of actual overlap, the values of the indices are distinct. An actual overlap which is measured by Czekanowski's Index as 0.5 varies between 0.32 and 0.71 depending on the index chosen (Fig. 1A).

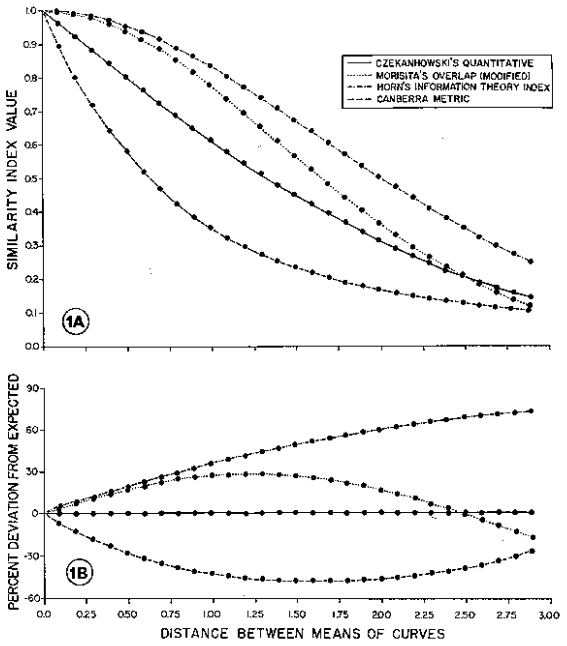


Fig. 1. Response curves of 4 similarity indices to decreasing overlap, based on normal distribution

The same point is made by Fig. 1B. An index that accurately reflects the true amount of overlap (as defined here) will show relatively little deviation from zero. Only Czekanowski's Index behaves in this manner. Horn's Information Theory progressively diverges and consistently overestimates similarity. Canberra Metric underestimates overlap for much of the range while Morisita's Index (modified) overestimates similarity over approximately the same portion of the range.

## DISCUSSION

There is a distinct danger of misinterpreting community analyses such as classifications and the significance of similarity values if the variance in answers introduced by the choice of a similarity index is not appreciated. A common and natural practice with

regard to similarity indices is to note the limits (0 to 1 for all indices examined here) and to divide mentally the range into equal intervals. One will then speak of very low, low, moderate, high, and very high similarities, based on a 0.2 interval. Unfortunately, what is 'high' to Horn's index would be 'moderate' to Czekanowski's and 'low' to Canberra Metric. Cross-study comparisons with different indices could be highly misleading and communication of results may suffer heavily from conversion to a potentially misleading qualitative scale.

While cluster-patterns in dendrograms are not radically affected by the non-linear correspondence of the indices to actual overlap, linkage values are affected. Relative to a dendrogram generated with Czekanowski's Index, Morisita's Index will contract linkages at high values while expanding the linkages for low similarity values. Clusters of high similarity will become more distinct and clusters of low similarity will approach zero, while intermediate links will become obscured. Conversely, Canberra Metric (which underestimates high values and overestimates low values) will tend to expand clusters of high similarity and contract clusters of low similarity. As a result, most of the links will lie close to the middle of the dendrogram, obscuring cluster relationships. Horn's Index consistently overestimates similarity and the dendrogram will tend to be shifted consistently to higher values.

One of the standard methods of reading dendrograms is to employ the 'fixed stopping rule' or to arbitrarily select a threshold similarity. If the linkage of a cluster is greater than that level, the cluster is regarded as important. Otherwise the cluster is ignored (Boesch, 1977). Obviously the choice of the index may then radically affect the number and identity of 'important clusters'. Even if a qualitative approach is taken in reading the dendrogram, the compression or expansion of clusters may be highly misleading.

The pattern shown in Fig. 1 points up a potentially critical problem. By selecting Horn's Index, a relatively great actual difference could occur between samples without being reflected by a commensurate change in index values. Conversely, the use of Canberra Metric would result in a greater change in index values than was actually justified. It is possible that either consciously or inadvertently, environmental impacts or experimental treatments could be over- or underestimated simply by the choice of a similarity index. For instance, changes in the community of a site undergoing or having undergone pollution stress may be viewed as minor (Horn's Index) or major (Canberra Metric), while the actual change was moderate (Czekanowski's Index). Unless a user is thoroughly familiar with this effect, major interpretive problems may result, and unless the scientific community appreciates the radical effect of index choice on the qualitative impression of community data (especially on non-technically trained persons), the potential for abuse exists.

The only index which accurately reflected similarity was Czekanowski's Quantitative Index. In that this index sums the lowest common value for overlapping blocks (species or resources), it is an analog to integration and can be expected to reflect actual overlap accurately for virtually any underlying distribution. This conclusion is predicated by the use of the area of overlap of two curves as being equivalent to 'true' similarity. The similarity is assumed to be symmetrical, e.g. the overlap of Curve A to B is the same as the overlap of Curve B to A. This conclusion holds for community studies but should be only cautiously applied to asymmetrical uses of similarity indices such as in niche overlap studies. I suggest that the term 'percent similarity' or 'percent overlap' be restricted to Czekanowski's Quantitative Index which in fact does measure that quantity.

Similarity indices have been used extensively in niche overlap (interspecific resource utilization) as well as in community similarity (overlapping density functions) studies. Many of the indices have firm theoretical and statistical foundations in the specific areas for which they were developed (Horn, 1966; Boesch, 1977). The discussion here is not aimed at niche studies (but see Hurlburt, 1978) but rather at the use of similarity indices in community studies by ecologists using computer programs without necessarily having extensively reviewed the pertinent literature. Care should be taken in interpreting and communicating the results of similarity indices and the justification for the use of a given index should be explicit.

## Appendix: Similarity Indices

Czekanowski's Quantitative Index (Bray and Curtis, 1957; Field and McFarlane, 1968):

$$CZ_{ik} = \frac{2\sum_{j=1}^{S} \min (x_{ij}, x_{kj})}{\sum_{j=1}^{S} (x_{ij} + x_{kj})}$$

Morisita's Index (modified by Horn; Horn, 1966):

$$M_{ik} = \frac{2\sum_{j=1}^{S} (x_{ij})\,(x_{kj})}{\sum_{j=1}^{S} x_{ij}^{2} + \sum_{j=1}^{S} x_{kj}^{2}}$$

Canberra Metric (Lance and Williams, 1967):

$$C_{ik} = \frac{\sum_{j=1}^{S} \left(\frac{(x_{ij} - x_{kj})}{(x_{ij} + x_{kj})}\right)}{S^{*}}$$

Horn's Information Theory (Horn, 1966):

$$R_{ik} = \frac{H_{max} - H_{obs}}{H_{max} - H_{min}}$$

where: $X = \sum_{j=1}^{S} x_{ij}$; $Y = \sum_{j=1}^{S} x_{kj}$; $H(X) = \sum_{j=1}^{S} \frac{x_{ij}}{X} \log \frac{X}{x_{ij}}$;

$$H(Y) = \sum_{j=1}^{S} \frac{x_{kj}}{Y} \log \frac{Y}{x_{kj}}$$

and

$$H_{max} = \sum_{j=1}^{S} \left( \frac{x_{ij}}{X+Y} \log \frac{X+Y}{x_{ij}} + \frac{x_{kj}}{X+Y} \log \frac{X+Y}{x_{kj}} \right);$$

$$H_{min} = \frac{X}{X+Y} H(X) \frac{Y}{X+Y} H(Y)$$

$$H_{obs} = \sum_{j=1}^{S} \frac{x_{ij} + x_{kj}}{X+Y} \log \frac{X+Y}{x_{ij} + x_{kj}}$$

In all equations, $x_{ij}$ = occurrence of the jth item (species or resource) in the ith sample (or consumer); $x_{kj}$ = the occurrence of the same item in the kth sample (or consumer); S = number of species or resources over all samples; $S^*$ = number of species actually present, e.g. joint absences are excluded.

*Acknowledgements.* I thank Drs. J. L. Simon, G. D. McCoy, B. C. Cowell, S. L. Santos and P. Feinsinger for critically reviewing this manuscript. This research was supported in part by NSF Grant GA-35120.

## LITERATURE CITED

Bloom, S. A., Santos, S. L., Field, J. G. (1977). A package of computer programs for benthic community analyses. Bull. mar. Sci. 27 (3): 577–580

Boesch, D. F. (1977). Application of numerical classification in ecological investigations of water pollution. Special Scientific Report 77, VIMS (EPA-600/3-7703)

Bray, J. R. Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconson. Ecol. Monog. 27 (4): 325–349

Dauer, D. M., Simon, J. L. (1975). Lateral or along-shore distribution of the polychaetous annelids of an intertidal, sandy habitat. Mar. Biol. 31: 363–370

Day, J. H., Field, J. G., Montgomery, M. (1971). The use of numerical methods to determine the distribution of the benthic fauna across the continental shelf of North Carolina. J. Anim. Ecol. 40: 93–126

Field, J. G. (1971). A numerical analysis of changes in the soft-bottom fauna along a transect across False Bay, South Africa. J. exp. mar. Biol. Ecol. 7: 214–244

Field, J. G., McFarlane, G. (1968). Numerical methods in marine ecology. 1. A quantitative 'similarity' analysis of rocky shore samples in False Bay, South Africa. Zool. Africana 3 (2): 119–137

Horn, H. S. (1977). Measurement of 'overlap' in comparative ecological studies. Am. Nat. 100 (914): 419–423

Hulburt, S. H. (1978). The measurement of niche overlap and some relatives. Ecology 59 (1): 67–77

Lance, G. N., Williams, W. T. (1967). Mixed-data classificatory programs. I. Agglomerative systems. Aust. Comput. J. 1: 15–20

May, R. M. (1975). Some notes on estimating the competition matrix, α. Ecology 56: 737–741

Rohlf, F. J., Sokol, R. R. (1969). Statistical tables, W. H. Freeman and Company, San Francisco, California

Santos, S. L., Simon, J. L. (1974). Distribution and abundance of the polychaetous annelids in a South Florida estuary. Bull. mar. Sci. 24: (3): 669–689

Williams, W. T., Lance, G. N., Webb, L. J., Tracey, J. G. (1973). Studies in the numerical analysis of complex rain-forest communities. VI. Models for the classification of quantitative data. J. Ecol. 61 (1): 47–70

This paper was submitted to the editor; it was accepted for printing on March 4, 1981