Ecology, 63(3), 1982, pp. 634–639

STATISTICAL SIGNIFICANCE OF SPECIES CLUSTERS IN ASSOCIATION ANALYSIS[1]

RICHARD E. STRAUSS[2]
Graduate Program in Ecology, Pennsylvania State University, University Park, Pennsylvania 16802 USA

Abstract. Cluster analysis techniques are often used in preliminary studies of community organization to identify groups of associated species on the basis of distributional co-occurrence. Adequate statistical tests have not been available for determining whether the associations recognized are sufficiently nonrandom to be considered significant, or might reasonably be expected on the basis of random distribution alone. For presence/absence occurrence data, an approximate test of significance may be performed by generating random occurrence matrices within the constraints of a randomization model which treats as marginal constants both the observed number of occurrences of each species and the observed number of species present at each locality. Null (random) distributions of node values of dendrograms, derived from the randomly generated occurrence matrices, are used to assign approximate critical test values. These values may be applied directly to the dendrogram derived from the observed distribution data to determine which of the observed species clusters may be considered statistically significant. Groups of significantly associated species are reasonable entities within which to examine ecological relationships further.

Key words: association analysis; cluster analysis; dendrograms; random distributions; randomization; species associations; species distributions.

INTRODUCTION

Cluster analysis techniques are regularly used in initial studies of community structure to attempt to delineate quantitatively distinctive associations of species within some specified geographic area. This is usually done with the hope of recognizing distinctive locality groups or plausible communities for further study. However, such studies rarely include an attempt to determine whether any of the groupings recognized are statistically significant in some sense, that is, whether the observed "associations" might not reasonably be expected on the basis of chance alone in the absence of biological interactions (Harper 1978, Connor and Simberloff 1978, 1979).

Harper (1978) has proposed an approximate statistical test for significant clustering by locality, based on a "Monte Carlo" randomization model, which has much potential for standard use in association studies. Its purpose is to test whether or not the overall degree of clustering of the locality data is statistically significant. It may also be applied to testing the significance of an overall clustering of species, and will be discussed here primarily in that context.

The purpose of this account is to show that Harper's test may be extended in two ways. (1) The randomization test model may be made more realistic by increasing the number of constraints, treating as fixed marginal constants both the number of occurrences of each species and the number of species at each locality. (2) Although the overall clustering pattern may be sufficiently nonrandom, not all of the individual groupings recognized may be statistically significant. The test may be altered to decide which, if any, of the observed clusters are significant.

THE STRUCTURE OF THE RANDOMIZATION MODEL

The randomization test proposed by Harper (1978) is designed for use with presence/absence data. It is based on a model of the results that should be expected if, within the constraints of the model, all elements are independent and varying at random. The elements in this case are the distributional occurrences of the observed species among the sampled localities. The raw data for the analysis are usually presented in the form of a matrix of N rows, one for each species, and M columns, one for each locality sampled. Each matrix element is 1 if the nth species was present at the mth locality, or 0 if it was absent. The occurrence matrix is converted to an $N \times N$ matrix of species-association indices by calculating pairwise measures of association for all possible combinations of species, using some suitable index of distributional association (Southwood 1968, Sneath and Sokal 1973). A dendrogram (clustering pattern) summarizing the overall structure of the distribution data may then be constructed from the association matrix (usually by first converting it to a matrix of "distance," or dissimilarity values) by applying one of a large number of available cluster analysis techniques (Sneath and Sokal 1973).

The null hypothesis to be tested is that the clustering pattern derived from the observed occurrence matrix does not differ significantly from what might be ex-

[1] Manuscript received 30 January 1981; accepted 10 June 1981; final version received 30 July 1981.

[2] Present address: Division of Fishes, Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109 USA.

pected by chance, based on the total population of possible occurrence matrices which could be generated by rearranging the 0's and 1's within the matrix. The general procedure for performing a randomization test for statistical significance involves (1) drawing random samples from the total population of possible configurations, (2) computing some test statistic for each random sample, (3) empirically determining the null frequency distribution of the test statistic, and (4) deciding on the significance of the original observation by comparing it to the empirical distribution. Harper recommends that the test for significant clustering be performed by randomly generating a sample of $N_t - 1$ occurrence matrices out of the total number possible, where N_t is the desired size of the random (null) frequency distribution to be used in the randomization test. For each matrix, including the one observed, some descriptive parameter R is calculated. The parameter that Harper proposes is the number of nodes of the derived dendrogram that occur below some arbitrary level of similarity c (such as 0.5), but any number of such descriptors might be appropriate. Regardless of how R is defined, the number of the N_t values of R which are greater than or equal to the R value obtained from the observed occurrence matrix can be calculated. If the proportion of this number out of the total N_t values is less than or equal to the desired level of significance α, then the null hypothesis can reasonably be rejected.

This basic randomization technique is a sound one, and in one form or another has had many applications in statistical analysis. In the present context it could be used with any random model, any clustering algorithm, any distance or similarity metric, and any descriptive parameter. However, the applicability and reliability of the method in any particular case will depend in large part on the assumptions and constraints of the model used to generate random matrices.

Although Harper does not explicitly discuss the assumptions underlying his random model, he recommends that random occurrence matrices be generated by rearranging the 0's and 1's in one or more rows of the observed matrix. Rearranging the elements in this way preserves the marginal row totals of the matrix, but the column totals become random variables. There are therefore two basic assumptions to the model as proposed. First, it is assumed that the relative frequency of each species among all localities is fixed at the observed value, and that the presence/absence data for each species are distributed among the M localities in proportion to the species' relative frequency. This is a realistic assumption and is a necessary attribute of the model although, to avoid circularity in the methodology, there should ideally be some independent means of determining the probability of a given species occurring at a particular number of sites (Connor and Simberloff 1978). In the absence of such data the observed frequency distribution may be used as an approximation of the abilities of species to disperse and persist throughout the area of study.

The second assumption of Harper's model is that for any species, the probability of occurring at a particular site is equal to that of occurring at any other site. This is equivalent to assuming that the species-carrying capacities of sites are equal, which might be true if the number of species at each locality were equal, or at least binomially distributed (Barton and David 1959). In most natural communities, however, different localities usually have different species-carrying capacities. For such situations the hypothesis that the number of species per locality has a binomial frequency distribution is so intrinsically unlikely as to be seldom worth testing (Pielou 1972). If Harper's test did indicate nonrandom structure in a particular set of species-occurrence data, it might be due to unequal species-carrying capacities among sites (i.e., to nonrandom structure in the column totals), to associations among species, or to a combination of both of these factors. Without further analysis there would be no way of knowing which would be the case. Thus the observed frequency distribution of the number of species per locality, in the absence of an independent assessment of the site-specific species-carrying capacities, is a valid and necessary set of information that should be preserved within the structure of the randomization test model (Connor and Simberloff 1979). Any nonrandom structure exhibited by the data when this known structure is taken into account should be due only to species association (that is, to statistical "interaction" between localities and species occurrences).

The deviation of an actual distribution of the number of species per locality from a binomial distribution may be illustrated with occurrence data on fish species in the Susquehanna River basin of Pennsylvania (Fig. 1A), representing 43 species and 642 collection localities. The expected binomial distribution, which obviously does not correspond to the observed frequency distribution, was calculated using the method of Barton and David (1959). The theoretical binomial can be closely approximated by a frequency distribution of the column (species per locality) totals of a random occurrence matrix generated with only the row (species occurrence) totals (Fig. 1B) held constant. Thus a very important aspect of the structure of the original occurrence matrix may be altered by generating a random matrix in this manner. The total numbers of species per locality should be treated as marginal constants if a significant portion of the observed data structure is not to be ignored in the analysis.

Testing for Statistically Significant Clusters

Harper's randomization test was designed for use in conjunction with a cluster analysis. As noted above,

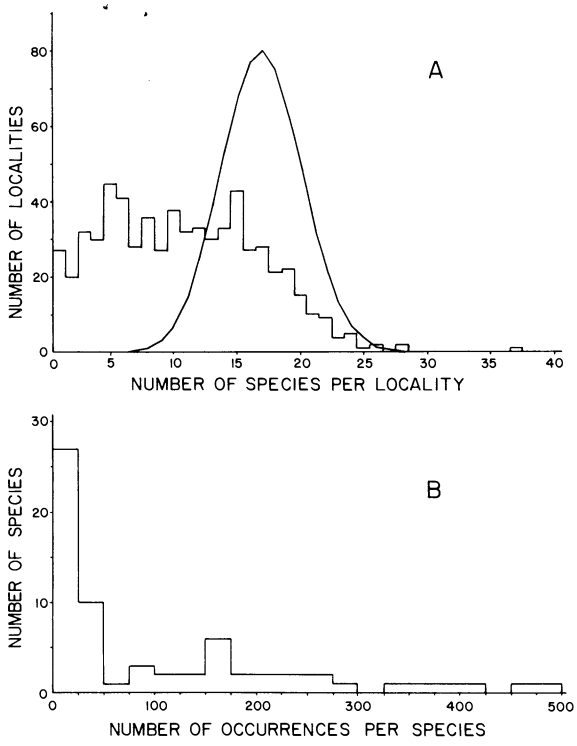

FIG. 1. Frequency histograms summarizing the occurrence of 43 species of fish at 642 collection localities throughout the Susquehanna River drainage of Pennsylvania. A. Histogram of the numbers of species collected at each locality, contrasted with the expected binomial distribution (continuous curve) of Barton and David (1959). B. Histogram of the frequencies of occurrence of the 43 species throughout the drainage.

its function is to test whether the overall degree of clustering of the dendrogram derived from the observed occurrence matrix is statistically significant with some chosen degree of confidence. But even though the overall clustering pattern may be sufficiently nonrandom to reject the null hypothesis of random association, not all of the individual clusters recognized may be significant. Various qualitative guidelines have been proposed for deciding at what point in the clustering process clusters become nonsignificant (e.g., Thorndike 1953, Marriot 1971, Mojena 1977). In general, however, and in ecological association studies in particular, the number of groups accepted by the investigator is usually chosen arbitrarily.

Everitt (1979), noting that the determination of the number of significant groups in a cluster analysis is a "formidable problem," has identified three principal difficulties encountered in deriving adequate significance tests: (1) specification of a suitable null hypothesis; (2) determination of the sampling distribution of the distance or similarity measure used; and (3) development of a flexible test procedure. With modification, the randomization test described above can be used to overcome these problems. A randomly generated sampling distribution of the measure of association being employed may serve as the basis for the null hypothesis of random association. No descriptive parameter of the sort used by Harper is necessary.

A test for significant association may proceed as follows. (1) Choose a level of significance α. For association analyses a level of $\alpha = .05$ is usually adequate. (2) Choose the number of association values (N_t) to be used in producing the null sampling distribution. The larger the value of N_t, the more resolution may be obtained in estimating the critical test value at significance level α. A reasonable lower limit for N_t is 1000. (3) Generate a random occurrence matrix, maintaining as fixed marginal totals the observed number of occurrences of each species and the number of species at each locality. (4) Calculate pairwise association values for all possible combinations of species (or localities, depending on the nature of the analysis) from the random matrix. A matrix for N species will yield $N(N - 1)/2$ association values. If the number of association values produced is less than N_t, repeat steps 3 and 4 until a sufficient number of values has been accumulated. If the chosen value of N_t is not at least several times larger than the number of pairwise association values derived from a single random matrix, then the matrix should be randomly subsampled. That is, to reduce the risk of generating an anomalous distribution, at least three or four random matrices should be used to produce the null distribution of association values. (5) Sort the association values into ascending sequence and determine the $1 - \alpha$ percentile. For example, if $\alpha = .05$ and $N_t = 1000$, the $1 - \alpha$ percentile would be the value of the 951st element in the sorted array. Any pairwise association value calculated from the original occurrence matrix may be considered statistically significant at the chosen level of significance if it exceeds this quantity. In the observed distribution of association index values for the fish species of the Susquehanna River basin (Fig. 2), 32.6% of all pairwise association values are greater than the critical value derived from the null distribution. Any of these values by itself may be considered to be statistically significant at the $\alpha = .05$ level of significance.

The $1 - \alpha$ percentile (or its equivalent) may be applied directly to the clustering structure of a dendrogram if some degree of caution is exercised. Because measures of dissimilarity rather than association are normally used to construct dendrograms (for example, by applying the 1's-complement of an index of association having a range of 0 to 1), a more appropriate critical test value would be the α percentile of a null distribution of the dissimilarity measure. Then any cluster could be considered conditionally significant if its node value is less than this α percentile. Caution is warranted by two factors: (1) the difference between the total null distribution of the dissimilarity (or as-

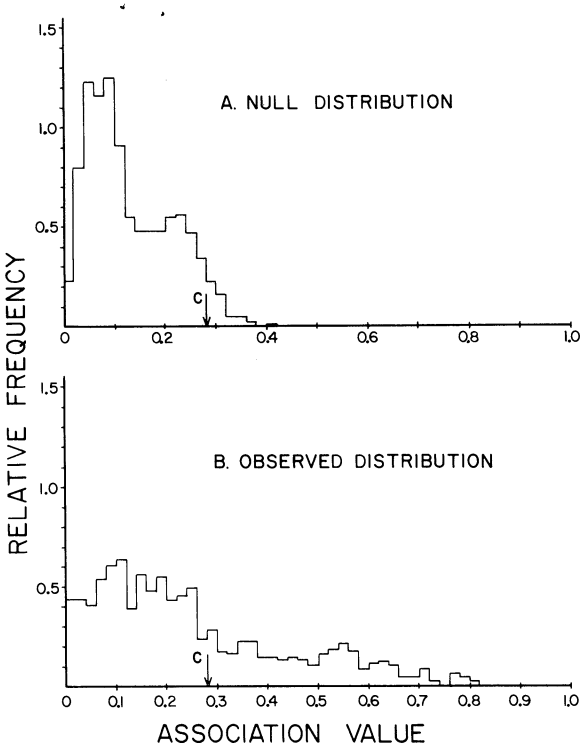

FIG. 2. Null and observed frequency histograms of pairwise association index values for the 43 fish species. The index used is Southwood's (1968) version of Fager's (1957) index of joint occurrence. A. Null distribution is derived from random occurrence matrices generated with fixed marginal totals. The histogram is based on 1000 index values. The arrow indicates the critical test value for $\alpha = .05$. B. Histogram of the 903 pairwise values calculated from the observed distributional data. The arrow indicates the critical value obtained from the null distribution.

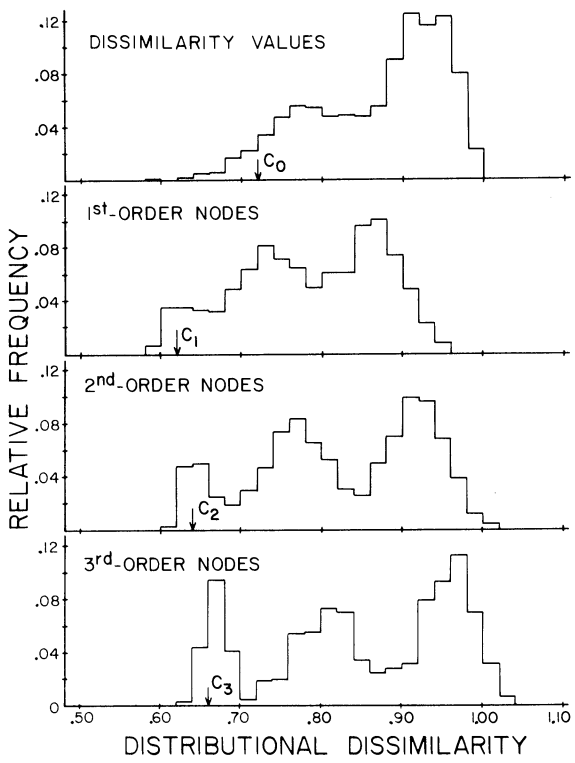

FIG. 3. Null frequency histograms of the dissimilarity index (the 1's-complement of the association index used in Fig. 2) and the first three node orders of derived dendrograms. The UPGMA clustering method (Sneath and Sokal 1973) was used to construct the dendrograms. Critical test values for $\alpha = .05$ are indicated by C_0 through C_3. Each histogram is based on 1000 randomization values.

sociation) measure and that portion of the distribution which actually appears in the low-order clusters of the dendrogram (a factor which will be discussed in more detail below); and (2) the distortion that appears within a dendrogram as the clustering proceeds to high-level groups, due to the representation of multidimensional distances in a space of progressively fewer dimensions. Different clustering algorithms may yield different amounts of distortion, which can be quantified by means of the cophenetic correlation coefficient of Sokal and Rohlf (1962). A high-distortion algorithm may give less statistical protection than expected; when the $1 - \alpha$ percentile is used to evaluate cluster nodes, the effective value of α will in this case actually be less than the value of α initially chosen. A conservative clustering algorithm may give more statistical protection than expected.

The exact statistical significance of cluster nodes may be estimated stochastically by modifying steps 4 and 5 of the above procedure in the following manner. (4) Calculate distributional dissimilarity values for all possible pairwise combinations of species from the random occurrence matrix that has just been generated, and then produce a dendrogram of these data. Accumulate the cluster node values for second-order clusters (linking a first-order cluster and a dissimilarity value, or two first-order clusters), third-order clusters (linking a second-order cluster and a dissimilarity value, a second-order and a first-order cluster, or two second-order clusters), and so on. Repeat step 3 and this modified step 4 until N_t values for each of the sets have been obtained. (5) Sort each set of values into ascending sequence and determine the α percentiles. The statistical significance of each nth-order cluster in the dendrogram derived from the observed data matrix can then be evaluated properly by referring the observed cluster node value to the corresponding α percentile.

The null distributions of the first three dendrogram node orders from a cluster analysis of the fish distribution data (Fig. 3) reveal two characteristics that are likely to be general properties of nodal distributions. The first is that the critical test value for each node increases with node order. This is a predictable con-

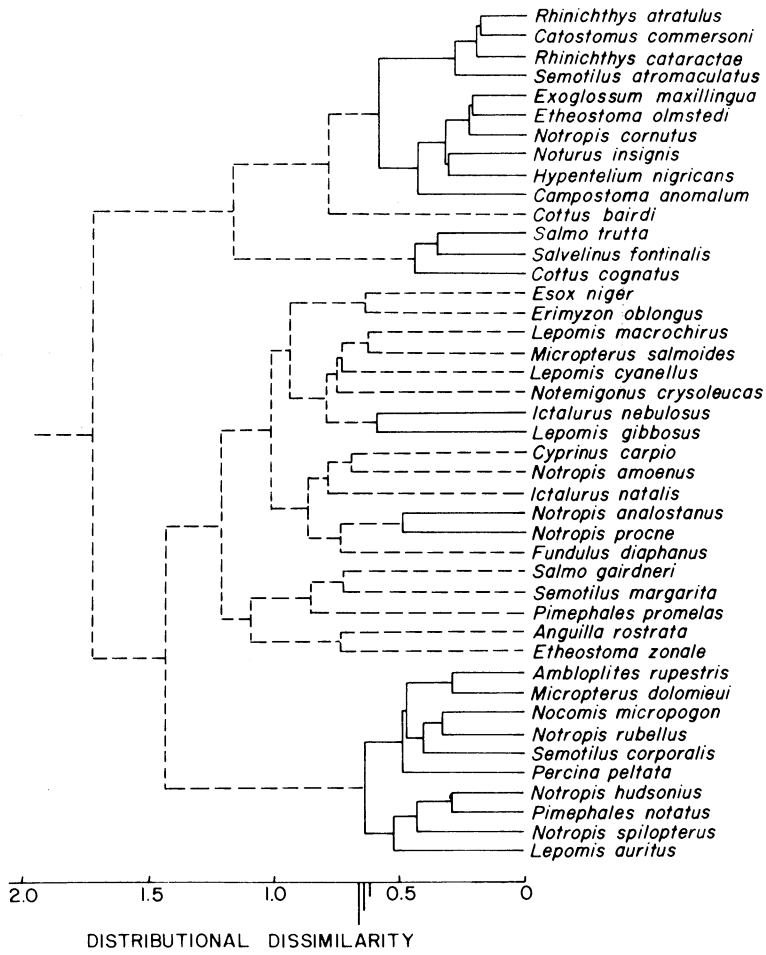

FIG. 4. Cluster analysis of the joint occurrence of 43 species of fish in the Susquehanna River drainage of Pennsylvania, constructed with the UPGMA clustering algorithm (Sneath and Sokal 1973). The three short perpendicular lines on the dissimilarity scale represent the critical values C_1, C_2, and C_3 obtained from the null nodal distributions of Fig. 3. Significant clusters are indicated by solid lines. The nonsignificant portion of the dendrogram is drawn in dotted lines.

sequence of the clustering technique itself, since high-order clusters are formed by combining lower order clusters. The second characteristic is that each of the critical node values is less than (and therefore more conservative than) the critical value of the original measure of dissimilarity (indicated as C_0 in Fig. 3). This is due to the fact that only a portion of the total null distribution of the dissimilarity measure is included in the distributions of low-order clusters; high dissimilarity values tend to enter as single elements into high-order clusters. Thus the application of separate critical test values for different levels of clustering is a more conservative and theoretically more valid procedure than the use of only a single critical value derived from the null distribution of the dissimilarity index.

Parenthetically, it should be noted that the particular method of assigning order rankings to cluster nodes which was described above in the modified step 4 is arbitrary and actually combines different topological classes of clusters into the same node orders. This is reflected in the polymodal nature of the nodal distributions of Fig. 3. A more refined scheme of assigning node-order rankings might give more exact estimates of statistical significance, but at the expense of increasing the number of null distributions to be generated and stored.

When critical values for the first three node orders are applied to the dendrogram derived from the observed fish distribution data (Fig. 4), three species clusters are clearly significant and several species pairs are marginally significant. The UPGMA clustering algorithm (Sneath and Sokal 1973) was used to produce the dendrogram, but any clustering technique might have been used. Of the three significant clusters, two represent associations of species characteristic of headwater streams while the third consists of species encountered in larger, lower gradient habitats. Such

associations may be of some interest in their own right as summary descriptions, but should preferably be used to formulate working hypotheses of ecological relationships to be tested experimentally.

The criterion of statistical association does not guarantee the existence of important interspecific interactions. Similarly, lack of significant association does not necessarily indicate absence of functional relationships, and might in fact be due to interspecific competitive interactions. Nevertheless, groups of associated species determined objectively from a preliminary analysis of distributional patterns do represent practical entities within which to examine ecological relationships further.

ACKNOWLEDGMENTS

I am grateful to E. L. Cooper for the use of his data on fish distributions in Pennsylvania. For comments and suggestions concerning the preparation of this manuscript I wish to thank E. L. Cooper, C. W. Harper, Jr., J. B. Horton, M. A. Houck Strauss, and J. L. Rosenberger. This work was supported by the Pennsylvania Agricultural Experiment Station and by National Science Foundation Grants DEB-7903285 and DEB-8011562. Authorized for publication as Paper 6014 in the Journal Series of The Pennsylvania Agricultural Experiment Station, University Park, Pennsylvania, USA.

LITERATURE CITED

Barton, D. E., and F. N. David. 1959. The dispersion of a number of species. Journal of the Royal Statistical Society, **B 21**:190–194.

Connor, E. F., and D. Simberloff. 1978. Species number and compositional similarity of the Galapagos flora and avifauna. Ecological Monographs **48**:219–248.

Connor, E. F., and D. Simberloff. 1979. The assemblage of species communities: chance or competition? Ecology **60**:1132–1140.

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

Fager, E. W. 1957. Determination and analysis of recurrent groups. Ecology **38**:586–595.

Harper, C. W., Jr. 1978. Groupings by locality in community ecology and paleoecology: tests of significance. Lethaia **11**:251–257.

Marriot, F. H. C. 1971. Practical problems in a method of cluster analysis. Biometrics **27**:501–514.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. Computer Journal **20**:359–363.

Pielou, E. C. 1972. Measurement of structure in animal communities. Pages 113–135 *in* J. A. Wiens, editor. Ecosystem structure and function. Proceedings of the 31st Annual Biology Colloquium. Oregon State University Press, Corvallis, Oregon, USA.

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

Sokal, R. R., and F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon **11**:33–40.

Southwood, T. R. E. 1968. Ecological methods. Methuen, London, England.

Thorndike, R. L. 1953. Who belongs in a family? Psychometrika **18**:267–276.