Vegetatio 72: 89–93, 1987

Affinity analysis: methodologies and statistical inference

Samuel M. Scheiner[1,2,3] & Conrad A. Istock[1,2]
[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 89721, USA*
[2]*University of Michigan, Biological Station, Pellston, MI 49769, USA*
[3]*Present address: Department of Biological Sciences, Northern Illinois University DeKalb, IL 60115, USA*

Accepted 28.5.1987

Keywords: Diversity, Meadow community, Mosaic diversity, Random simulation, Similarity

Abstract

Affinity analysis (AA) is a group of methods for the study of the variation of degrees of compositional relatedness among all of the communities in a landscape. AA can be used with statistical inference to compare mosaic diversities ($\overset{*}{\mu}$) among different landscapes, identify unusual sites within a landscape, and determine when a pair of sites are significantly different from each other. These procedures were done with a set of samples from meadows in the Danube River Valley near Ulm by using random simulations to provide expectations of the summary statistics. The sampling and statistical limitations of AA were discussed.

Introduction

The assessment of pattern and the measurement of diversity within and among communities is a central focus of ecology. Of special interest is the recent debate over the existence and cause of patterns in natural communities (Connell 1983; Roughgarden 1983; Quinn & Durham 1983; Simberloff 1983; Strong 1983; Noy-Meir & Van der Maarel 1987). Older concerns over methods for the measurement of diversity at different levels of organization (Whittaker 1972; Peet 1974; Wilson & Shmida 1984) also remain worthy of extension (Pielou 1975). As the method used for analysis can effect the conclusions reached (Peet 1974; Diamond & Gilpin 1982; Harvey *et al.* 1983; Wilson & Shmida 1984), it is important to establish techniques whose assumptions are clear and whose statistical properties are understood.

Affinity analysis (AA) provides methods for the study of varying degrees of compositional relatedness among the communities in a landscape (Istock & Scheiner 1987). AA provides a new measure of high-order diversity of the landscape mosaic. In addition, AA allows statistical inference in the comparison of different samples at three levels of analysis. First, by providing a measure of the diversity of communities in an entire landscape, AA can be used 1. to determine if a landscape mosaic is more diverse than expected at random and 2. to compare the diversities of different landscapes. Second, through calculation of the affinity of each community to the landscape, AA can be used to find those communities which represent large departures from the central tendency of the landscape. Finally, the results of AA can be used to decide when any pair of sites are significantly different from each other with respect to either pairwise similarity or affinity.

Affinity analysis consists of two preliminary steps. First, the original data matrix, consisting of sites (columns) by species (rows), is transformed to a site-by-site matrix of similarity coefficients. The mean similarity of each site ($\bar{S}_i$) is calculated. A second transformation changes the similarity matrix into a site-by-site matrix of pairwise, signed, Wilcox-

on T values. These values are the relative affinities of the two sites to the rest of the landscape. The mean affinity of each site ($\bar{T}_i$) is computed. The mosaic (μ) diversity of the landscape as a whole is computed by taking the slope of the relation between mean similarities and mean affinities of the sites, the $S-T$ relation.

Computer simulations

Three parameters will describe any presence-absence data set: the number of sites (Q), the total number of species (R), and the fraction of entries which are represented by a 1 (matrix filling, F). For any combination of parameters (Q, R, and F) random data sets can be constructed as follows. For each entry in the data matrix a random number from 0 to 1.0 is chosen from a uniform distribution. If the random number is smaller than the matrix filling, the entry is assigned a value of 1, otherwise it is assigned a value of zero. This procedure does not constrain either the row or column totals. It is equivalent to a null hypothesis that all species have equiprobability of being in any site, i.e., that the species are randomly distributed throughout the simulated data matrix. We agree with critics of this null hypothesis (e.g., Diamond & Gilpin 1982, see Noy-Meir & Van der Maarel 1987) that these conditions will rarely be met in nature. However, as with standard assumptions of statistics such as random, independent, and normal distributions of observations, this assumption provides a baseline of variation to which natural variation can be compared.

The random matrix is then analysed using affinity analysis and summary statistics extracted. These statistics are the mean Jaccard similarity of all sites $\bar{\bar{S}}$ (Jaccard 1901), the variance in similarity among all sites $V(S_{ij})$, the variance in mean similarity of individual sites $V(\bar{S}_i)$, the variance among all sites of the Wilcoxon T statistics $V(T_{ij})$, the variance in mean T of individual sites $V(\bar{T}_i)$, and mosaic diversity (μ). This procedure is repeated thirty times and the random expectation and variance of each of the summary statistics is computed. Preliminary analysis indicated that within thirty runs the coefficients of variation of $\bar{\bar{S}}$ and μ ceased to change. This procedure is formally equivalent to a bootstrap (Efron 1981).

In a similar fashion random expectations can also be obtained for abundance data. The equivalent procedure with abundance data is done by filling each entry in the random data matrix with a randomly chosen entry from the original data matrix being tested. Sampling is done with replacement. Again, row and column totals are not constrained. This procedure is equivalent to constructing an average abundance distribution based on the original data and sampling all species in all sites from that distribution. Again, this null model is primarily designed to provide a baseline of variation and is not meant to mirror reality.

In order to understand the behavior of the AA parameters and develop empirical rules for its use we repeated the above analysis over a range of primary parameter values. The number of sites ranged from 10 to 100. The total number of species ranged from 10 to 1600. Matrix filling ranged from 0.05 to 0.90. These dimensions were chosen because they encompass most data sets of natural vegetation. In addition to the means and variances of the parameters we calculated skews and the correlation between $\bar{\bar{S}}$ and μ. Copies of the computer programs in BASIC and FORTRAN to perform affinity analysis and the bootstrapping procedure are available from us.

Statistical inference

The first step, before performing any statistical tests, is an assurance that the parameters to be tested are well behaved. The two measures of diversity, $\bar{\bar{S}}$ and μ as well as the other statistics, were found to vary in regular fashion as a function of the three primary data matrix parameters over most of the range of values tested. The diversity indices were not correlated with each other. The coefficients of variation of all of the parameters were small (0.1–0.3) and little skew was found.

Statistical inference can be accomplished at three levels of comparison: the whole landscape to its null expectation, single sites to the landscape, and between individual sites. These procedures will be illustrated with presence-absence data of a set of samples

from meadows on the Danube River near Ulm, southern Germany (Mueller-Dombois & Ellenberg 1974).

At the level of the whole landscape two types of comparisons can be made. First, μ can be used to determine if a landscape mosaic is more diverse than expected at random. From the random simulation we found $CV(\mu) = \sigma\mu/E(\mu) < 0.3$, or:

$$\sigma\mu < 0.3E(\mu). \quad (1)$$

In Istock & Scheiner (1987) we defined the proportional deviaton from the random expectation of μ to be:

$$\overset{*}{\mu} = [\text{Obs}(\mu) - E(\mu)]/E(\mu). \quad (2)$$

The departure, in units of standard deviations, of a sample from a theoretical distribution is given by the z statistic (Snedecor & Cochran 1967):

$$z = [X-E]/\sigma, \quad (3)$$

where X is the measured quantity and E and σ are the theoretical mean and standard deviation, respectively. On substituting we obtain:

$$z > \overset{*}{\mu}/0.3. \quad (4)$$

Because the distribution of μ is somewhat skew we recommend using three standard deviations as a conservative test of whether the mosaic diversity of a landscape is significantly different from random expectation.

In the analysis of the Danube River meadow data we found that $\mu = 0.00161$ and $E(\mu) = 0.0058$ giving $\overset{*}{\mu} = 1.78$. So, by Eq. 4, the Danube meadow samples have a mosaic diversity six standard deviations ($P < 0.00001$) greater than random expectation.

Comparisons of mosaic diversities among different landscapes can also be done. For this comparison we use $\overset{*}{\mu}$ in order to correct for differences in the values of the primary data matrix parameters. Istock & Scheiner (1987) presented a comparison of seven field data sets and found that $\overset{*}{\mu}$ tended not to vary greatly among landscapes. Although presently we are not able to do so, statistical inference can be done with these comparisons once an adequate sample size is reached to provide a good estimate of the expected mean and variance of $\overset{*}{\mu}$ in natural data sets.

Within a single landscape sites which are significantly different from the mode can be identified. If a site has a $\bar{T}_i$ that is several standard deviations, as given by the square root of the measured value of $V(\bar{T}_i)$, away from the mode that site represents a community quite different from the central tendency of the landscape. The T axis in the $S-T$ graph (Fig. 1) could be scaled in standard deviation units so as to display these differences directly. For example, in the Danube meadow data set, $V(\bar{T}_i) = 752$ and $\bar{T}_4 = 51.1$, which is 3.3 standard deviations away from modal sites 5 and 13.

Affinity analysis presents two different measures for the comparison of individual sites. First, the pairwise similarity values indicate which sites are more similar than would be expected at random.

$$z = [S_{xy} - E(\bar{\bar{S}})]/\sqrt{E(\bar{\bar{V}}(S_{ij}))}, \quad (5)$$

where S_{xy} is the measured similarity of any sites x and y, and $E(\bar{\bar{S}})$ and $E(V(S_{ij}))$ are the random expectation of the mean and variance of the similarity S_{ij} for any pair of sites obtained from the bootstrap. For eaxmple, in the Danube meadow samples $E(\bar{\bar{S}}) = 0.19$ and $E(V(S_{ij})) = 0.00363$. For sites 4 and 11, $S_{4,11} = 0.25$ and by Eq. 5, $z = 1.00$, $P < 0.16$. So these two sites are no more similar than would be expected at random. In contrast,

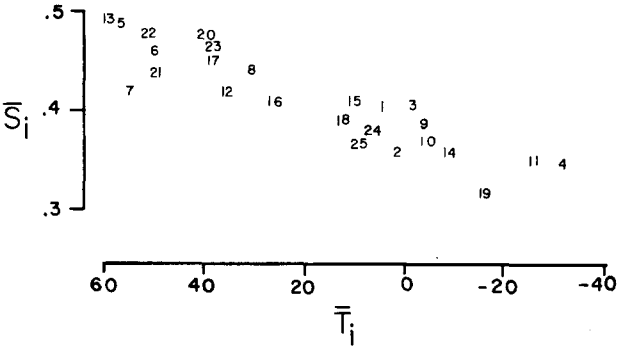

Fig. 1. Affinity analysis: $S-T$ graph of a set of meadow sites in the Danube River valley (Mueller-Dombois & Ellenberg, 1974). Site numbers as in the original data. Primary data dimensions are $Q = 25$, $R = 94$, and $F = 0.34$, $\bar{\bar{S}} = 0.41$ and $\mu = 0.00161$.

$S_{23,24} = 0.51$ which yields $z = 5.31$, $P < 0.00001$. So sites 23 and 24 are more alike than would be expected at random.

This test of similarity is complementary to that presented by Janson & Vegelius (1981). They present a formula for the standard error of the Jaccard similarity index which can be used to determine when two sites are more dissimilar than expected at random.

A second comparison among individual sites using the Wilcoxon T statistic provides information distinct from that carried solely by the similarity indices. The pairwise T values indicate which sites have different affinities to the landscape as a whole. In this case the theoretical expectation for T is given by the formula:

$$E(T) = N(N+1)/4 \text{ (Siegel 1956),} \quad (6)$$

where $N = Q-2$, and:

$$z = [\,|T_{xy}| - E(T)]/\sqrt{E(V(T_{ij}))}, \quad (7)$$

where $|T_{xy}|$ is the absolute value of the Wilcoxon T for sites x and y and $E(V(T_{ij}))$ is the random expectation of the variance as calculated from the bootstrap. (The theoretical expectation of $V(T_{ij})$ given by Siegel (1956) is based on the total sum of ranks, not the smaller sum of ranks as is used here, and so is not appropriate for Eq. 7.) For example, in the Danube meadow samples, $|T_{4,11}| = 128$ and $z = -0.32$, $P < 0.37$ while $|T_{23,24}| = 20$ and $z = -3.82$, $P < 0.0001$. So, sites 4 and 11 have equal affinities to the landscape while sites 23 and 24 have different affinities. In general, there is only a loose concordance between pairwise comparisons based on the similarity index and the T statistic. If two sites are identical, or nearly so, then perforce they must have equal affinities with the landscape. But as the pairwise similarity decreases, depending on the exact identity of the species *not* shared by the two sites their affinities may or may not diverge. And two sites with few or no species in common may still have similar affinities with the rest of the sites (e.g., sites 4 and 11, Fig. 1).

Empirical rules

The random simulations provide a set of empirical rules for the use of affinity analysis.

1. At extreme values of matrix filling (F), the method can not resolve structure because there is little or no variation among sites as expressed by variation in species composition. As seen in Fig. 2, at very low (< 0.10) and very high (> 0.80) values of F, the variance in $\bar{T}_i$ approaches zero. At low F (i.e., low mean similarity) AA now indicates that there is virtually no resemblance among the sites; and at high F, AA indicates that all the sites are nearly identical. Thus at the low end there is no sense of continuity in the data, the landscape is exceptionally fragmented, and at the high end there is no variation among the sites left to measure.
2. If the number of sites (Q) is small (< 20), we found that sampling errors make it difficult to separate structure from random noise. It would also, on ecological grounds, often be difficult to characterize a landscape on as few as 10 to 15 sites.

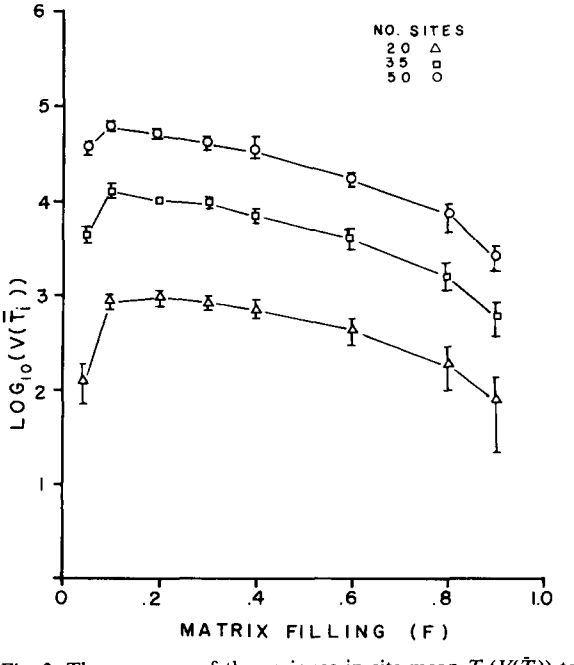

Fig. 2. The response of the variance in site mean T ($V(\bar{T}_i)$) to changes in matrix filling (F) for various numbers of sites. The number of species is 8 times the number of sites.

3. If the number of species (R) is less than three times the number of sites, the method encounters difficulty assessing structure. Species provide the basis for affinities in the landscape through their distribution patterns. As each species is scattered upon the landscape it provides an independent assessment of the many environments (Gleason, 1926). Since the number of environmental combinations likely increases with each additional site, more species are necessary to provide an adequate analysis. Obviously, in any real data set the three parameters (Q, R, and F) will interact. As more sites – especially more divergent ones – are sampled, more species are encountered. But, as the sites become more divergent, the matrix filling decreases. To increase the matrix filling, we suggest, for presence-absence data at least, that sites which seem almost redundant in species composition be included in any sampling design. The inclusion of subtle variation in composition will more clearly define and order the whole range of affinities.

Acknowledgements

We thank S. Naidu for programming assistance and R. Caplan for Fig. 1. This work was supported, in part, by the Naturalist Ecologist Training Program of the University of Michigan Biological Station and Ms. Judy Scheiner.

References

Connell, J. H., 1983. On the prevalence and relative importance of interspecific competition: Evidence from field experiments. Am. Nat. 122: 661–696.

Diamond, J. M. & Gilpin, M. E., 1982. Examination of the null model of Conner and Simberloff for species co-occurrences on islands. Oecologia 52: 64–74.

Efron, B., 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68: 589–599.

Gleason, H. A., 1926. The individualistic concept of the plant association. Bull. Torrey Bot. Club. 53: 7–26.

Harvey, P. H., Colwell, R. K., Silvertown, J. W. & May, R. M., 1983. Null models in ecology. Ann. Rev. Ecol. Syst. 14: 189–212.

Istock, C. A. & Scheiner, S. M., 1987. Affinities and high-order diversity within arrays of communities. Evolutionary Ecology 1: 11–29.

Jaccard, P, 1901. Distribution de la flore alpine dans le Bassin des Dranes et dans quelques regions voisines. Bull. Soc. Vaud. Sci. Nat. 37: 241–272.

Janson, S. & Vegelius, J., 1981. Measures of ecological association. Oecologia 49: 371–376.

Mueller-Dombois, D. & Ellenberg, H., 1974. Aims and methods of vegetation ecology. John Wiley & Sons, New York.

Noy-Meir, I. & Van der Maarel, E., 1987. Relations between community theory and community analysis in vegetation science: some historical perspectives. Vegetation 69:.

Peet, R. K., 1974. The measurement of species diversity. Ann. Rev. Ecol. Syst. 5: 285–307.

Pielou, E. C., 1975. Ecological diversity. John Wiley & Sons, New York.

Quinn, J. F. & Dunham, A. E., 1983. On hypothesis testing in ecology and evolution. Am. Nat. 122: 602–617.

Roughgarden, J., 1983. Competition and theory in community ecology. Am. Nat. 122: 583–601.

Siegel, S., 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York.

Simberloff, D., 1983. Competition theory, hypothesis testing, and other community ecology buzzwords. Am. Nat. 122: 626–635.

Snedecor, G. W. & Cochran, W. G., 1967. Statistical methods. The Iowa State University Press, Ames, Iowa.

Strong Jr., D. R., 1983. Natural variability and the manifold mechanisms of ecological communities. Am. Nat. 122: 636–660.

Whittaker, R. H., 1972. Evolution and measurement of species diversity. Taxon 21: 213–251.

Wilson, M. V. & Shmida, A., 1984. Measuring beta diversity with presence-absence data. J. Ecol. 72: 1055–1064.