A SYNOPSIS OF ECOLOGICAL CLASSIFICATION NUMERICAL METHODS

David T. Cleland and Carl W. Ramm

Great Lakes Ecological Assessment White Paper

A SYNOPSIS OF ECOLOGICAL CLASSIFICATION NUMERICAL METHODS

David T. Cleland and Carl W. Ramm[1]

Ecological classifications in various forms have been developed using relevees (species lists ordered by sampling quadrants), tabular analyses, and more recently by various multivariate numerical methods (see Host and Pregitzer 1991, Retuerto and Carballeira 1991, Spies and Barnes 1985a, Gauch 1984, Mueller-Dombois and Ellenberg 1974). Braun-Blanquet's procedure of 1921 used tabular methods in successive approximations to identify groups of species occurring in similar samples, and to identify samples with similar species composition. These early tabular classification techniques were informal and inherently subjective (Whittaker 1960, Mueller-Dombois and Ellenberg 1974), and thus the recognition of differential species groups and groups of similar samples depended on the individual investigator's understanding of species-species and species-environmental relationships within a study area.

In the past few decades, more objective multivariate procedures have been applied in ecological classification research (Digby and Kempton 1987). Research problems are often approached in an overall program using ordination for indirect and direct gradient analyses, clustering to group samples and variables, and tabular synthesis of results to corroborate patterns detected through ordination and clustering (Gauch 1982).

In these procedures, explorative data analyses are used to detect intercorrelations among variables, to check assumptions of data structure underlying particular analyses and suggest appropriate transformations, and to identify sample outliers using descriptive statistics and graphical displays of raw data. Ordination is also used in exploratory data analysis to detect patterns and outliers, and in multivariate analysis to screen variables, reduce dimensionality, and summarize community and environmental patterns (Gauch 1982).

Ordination is often followed by cluster analysis to identify ("recover") natural groups of samples and species. Results of ordination and clustering may be compared, and subsets of data may be interrogated to further elucidate relationships. Several complementary analysis techniques may be applied to the same data set, with the analysis progressing by successive refinement. Community patterns may be compared with environmental patterns to produce an integrated interpretation of ordination and classification results. The communication of results is promoted, however, by employing a moderate number

[1] Ecologist, North Central Research Station, USDA Forest Service, Rhinelander Wisconsin; and Associate Professor Forest Biometry, Department of Forestry, Michigan State University, East Lansing, MI.

of commonly used, relatively standardized methods (Pielou 1984, Romesburg 1979). Finally, hypothesis testing methods may be used a posteriori to assess ecosystem-level differences in processes such as productivity (Host et al. 1988), successional pathways (Host et al. 1987) and nutrient cycling (Zak et al. 1986, 1990).

Ordination

The goal of ordination is typically to place the field samples in a sequence such that their order reveals any underlying environmental or ecological gradient(s). Ordination, a primary tool for examining continuous change in ground-flora distributions, is a matrix approximation technique used to summarize data in a scatter diagram (Bray 1957), and reduce the dimensionality of a multivariate data set (Causton 1988, Gauch 1982). The reduction of dimensionality results from finding the "best" sub-space projection of the original n points in p-dimensional space to k dimensions ($k<n, p$). (Morzuch and Ruark 1991, Krzanowski 1988). In ecological studies, ordination is also used to discover latent structure of vegetation data due to species' responses to underlying environmental gradients (Prentice 1977).

Ordination uses spectral decomposition or singular value decomposition to build new coordinates for the data. The decomposition used in these methods is based on an assumed underlying linear model (Morrison 1976). Principal component analysis (PCA; Morrison 1976), correspondence analysis (CA; Hill 1974, Greenacre 1984, 1993), and detrended correspondence analysis (DCA; Hill and Gauch 1980) are among the most commonly used ordination techniques in modern ecological studies. These are variance-maximization methods that involve rotating the axes of a multidimensional cloud of points in multivariate space. This maximization of variance is equivalent to minimizing the variance of the projection distances from the axis. The first axis accounts for the majority of the variance in the ordination. A second axis is then found which is orthogonal to the first axis and that accounts for the maximal remaining variance, and so on, for as many axes as desired.

Geometrically, the general intentions of PCA, CA, and DCA are identical; a multidimensional cloud of points is projected efficiently in fewer dimensions which maximally account for the structure of a cloud of points in multidimensional space (Oksanen and Huttunen 1989, Gauch 1982). The methods of projection differ, however, due to the selection of particular distance measures, selection of the weights attributed to the points, and the positioning of the origin (Greenacre 1984). Centered PCA uses Euclidean distances, equal weights for points, and location of the origin at the centroid; non-centered options are also available. CA and DCA use chi-squared distances, weights for sample or species points proportional to the total for the species or sample, respectively, and an origin at the centroid (Liebart et al. 1984).

Principal Component Analysis

PCA may be used to reduce the dimensionality of a data set with a relatively large number of correlated variables. PCA creates new uncorrelated variables that are linear combinations of the original data (Newcomer 1984). The new variables retain almost all of the information content of the original data, but in a reduced subspace (Morzuch and Ruark 1991, Jongman et al. 1987). The first component accounts for the majority of the generalized variance present in the original data set; successive components account for decreasing proportions of the variance while remaining uncorrelated with previous components.

The relative amount of the total variation explained by each principal component is calculated by examining its variance (i.e., the eigen roots) relative to the overall variability – the sum of the eigen values (Morrison 1976). If a large proportion of the overall variability in a data set is accounted for in the first few principal components, these components can be used to summarize the variability of the original data set (Morzuch and Ruark 1991). In the end, however, PCA results must be assessed in terms of ecological utility; mere percentage of variance accounted for has not been found a reliable indicator of the quality of results (Gauch 1982).

PCA requires continuous data as it operates on a variance-covariance matrix or correlation matrix. The correlation matrix is commonly used when variables have dissimilar scales and ranges or have unequal variances (Padley 1989). Data with dissimilar scales and unequal variances will inevitably emphasize variables with large means and variances, and such emphasis may be unwarranted. The use of the correlation matrix avoids rather than solving this problem, but is considered satisfactory if all the variables used are of similar importance (Chatfield and Collins 1980, Morrison 1976).

Examining the coefficients of the eigen vectors associated with each PC permits identification of the variables that are most important to each PC. The set of variables so identified are commonly used in predictive models (Padley 1989, Pregitzer and Barnes 1984). Important variables may also be identified by examining the correlations, also called factor loadings or factor patterns, of the raw variables with the PC's; factor loadings should correspond with factor weights. Variables with high correlations are those most strongly associated with the respective dimension being considered.

Correspondence Analysis

Correspondence analysis (CA) was developed to analyze counts or measures of species abundance (Causton 1988, Host 1987). CA works with an n by p matrix of counts to produce an ordination of the p species and of the n samples. Species scores, when graphed, reflect the relative positions of individual species across all samples in response to one or more environmental or temporal (e.g., disturbance)

gradients (Hauser and Mucina 1981). Sample scores can be used to produce another ordination, with similar samples plotting close together and dissimilar samples plotting far apart. CA also provides information on the importance of individual species and sample points, and the relationship of species and samples to principal axes, termed coordinates, in a reduced subspace using a geometric interpretation (Greenacre 1984).

A cross-tabulation is converted into a contingency table in CA by dividing profiles of variable and sample frequencies by variable and sample totals. These sets of relative frequencies add up to a constant of one or 100, depending on whether the data are proportions or percentages, respectively. The geometry of such data has special properties, conferred by properties of a reference circle with a radius of one, that allow a geometric interpretation of the data structure. Rather than simply representing species in sample space, or samples in species space, a vector going from the origin to the sample points could represent samples or species (Greenacre 1993, 1984). It is then possible to describe much of the information about sample or species relationships in terms of the angles between pairs of vectors, because the angles can be expressed by their cosines. This has statistical meaning because the cosine is identical to the correlation coefficient between the two samples or species. Cosines of angles between vectors for sample pairs are analogous to a samples-by-samples dissimilarity matrix for which the dissimilarity measure used is the correlation coefficient after subtraction from one (Greenacre 1993).

CA uses the chi-squared statistic as a measure of the discrepancy between the observed frequencies in a contingency table and the expected frequencies. A scaled chi-squared distance metric can be used to evaluate how much row or column profiles differ from their average profiles. The chi-squared distance function is similar to the Euclidean distance in physical space, except that each squared difference between coordinates is divided by the corresponding element of the average profile (Liebart et al. 1984).

In a CA, each row or column profile of relative frequencies has a unique weight associated with it, called a mass, which is proportional to the column or row sum, respectively, in the original cross-tabulation. The average profile is the sum of each profile weighted by that profile's mass.

The total variability, or inertia, is the chi-squared statistic divided by the grand total of the table. Total inertia can be decomposed into principal inertias associated with each principal axis, into inertia associated with each row (and column) of the data matrix, and into the partial contributions of each element to each respective principal inertia (Greenacre 1993).

That component of inertia along a principal axis, called principal inertia, is the inertia of the row points (or column points) projected onto the axis (Greenacre 1993, 1984). Hence the first principal inertia is the total inertia of a set of

projected points lying on the straight line along the first principal coordinate. Second and higher axis principal inertias are determined using the same process.

Principal inertia can be further broken down into row and column components along individual principal axes, termed partial contributions (Greenacre 1993, 1984). These partial contributions to inertia for either sample or species sum to one in each dimension, and are a measure of the contribution of each entity to the inertia (variance represented) of that principal axis.

The relative contributions of each principal axis to a point's inertia are also determined in CA. In the geometric interpretation of these relative contributions, the cosine of an angle between a given point in multidimensional space is determined by projecting that point perpendicularly onto the principal axis. A vector is then extended to a point intersecting a circle in standard position with a circle of radius one. This point has a pair of coordinates that can be described by cosine and sine functions; the cosine of the angle is equal to the ratio of the derived vectors. The square of this ratio is the contribution of the respective principal axis to the sample or variable inertia. Relative contributions are equivalent to squared correlations between the profile points and the principal axes (Greenacre 1993, 1984).

The decomposition of each principal inertia over the samples or rows is a measure of the contributions of each principal coordinate to the inertia of these profiles. This decomposition indicates which points are best explained by the combination of axes forming subspaces by examining the squared cosines for each axis (Greenacre 1993, 1984).

A particularly useful feature of CA is the geometric interpretation of the quality of individual points in a given subspace (Greenacre 1984). The squared cosines of the angles between a point and each axis of a set of axes can be added across any number of dimensions; the sum of squared cosines across all dimension sums to one. This sum of squared cosines, or quality, is a measure of how well points are represented in a particular set of axes. An examination of the quality of display of each point helps to diagnose which points are far off the plane and whose positions are thus inaccurate in the map. The accuracy of display, or display quality, is the percentage of inertia accounted for in a specific subspace (Greenacre 1993, 1984).

Points are often distributed along an arch in CA. Some points, however, may be interior, or even central to the arch, and other points may be on the periphery of the arch. Interior points indicate species with especially broad or undiscriminating (weedy) distributions, and similarly, samples with a broad or mixed collection of species. Peripheral points indicate species with especially narrow distributions, having strong indicator value, and suggest samples with a simple collection of similar species (Gauch 1982).

Detrended Correspondence Analysis

In vegetation analyses, the use of PCA and CA depends upon the assumption that species abundance changes linearly along environmental gradients. When nonlinearities occur, CA produces an arch effect due to the compression of first axis (Peet et al. 1987, Hill and Gauch 1980). Moreover, a vegetation data matrix generally consists of some measure of species presence or abundance arranged by species and samples. Since species are often absent from many sample plots, data matrixes often contain a high proportion of zeros. For reasons inherent to the mathematics of the technique, these high proportions of zeros also tend to distort the second and subsequent principle axes (Host and Pregitzer 1991). If the data matrix contains a high proportion of zeros, then PCA using the correlation matrix should also be avoided.

Hill (1979) argues that the arch effect is a high-order dependence of the second and higher dimensions on the first dimension, so that a given distance of separation in the ordination does not carry a consistent meaning in terms of implied differences between the samples or species. If these derived orthogonal and linearly independent axes are related by a quadratic or higher-order relation, information on important secondary gradients in the data may be deferred to higher axes. The interpretation of results is thus made more difficult because spurious axes must be distinguished from valid axes, and higher axes in higher dimensions may have to be explored to detect meaningful gradients (Hill and Gauch 1980).

This limitation of traditional ordination methods led to the development of DCA (Hill 1979). The detrending procedure aims to remove the quadratic dependence of the second axis on the first axis while extracting a second axis. In detrending, the first axis is divided into segments, and within each segment, the sample scores are readjusted to have a zero average. The readjustment of scores results in a set of detrended scores that are used to derive a new axis based on residual variation.

Wartenberg et al. (1987) are skeptical of the value of detrending. They assert that the archlike curvature is an inherent property of successive-replacement data that results from the partial overlap in the distribution of species along a one-dimensional environmental gradient. They suggest that DCA can hide the real data structure or even introduce new distortions. Similarly, Greenacre (1984) criticized detrending because the control over geometry is lost. Okasen (1988) found that when the two first eigenvalues of CA are close to each other, their order can be reversed due to random variation in the data. In DCA, the second axis is detrended with respect to the first and therefore very variable configurations result when the orientation of the first axis in the plane is changed. This can lead to situations where the detrended solutions are very unstable under random variation and therefore can only be casually interpreted. These findings suggest additional limitations in the use of DCA.

Van Groenewoud (1992) recommended that the only way to get reasonable results using CA or DCA is to restrict the sampling mainly to one gradient, or pre-stratify the samples before analysis to represent mainly one gradient at a time. Concluding that all ordination methods are influenced by data curvature and scaling, Greenacre (1993) recommends reporting the arch unscaled in two dimensions.

Clustering

The objective of clustering procedures, commonly called classification, is to identify naturally occurring groups based on all variables in a multivariate data set (Host et al. 1993). Both the process of classification and the choice among classification techniques are more complex and more subjective than those of ordination (Gauch 1982). The utility of a given technique is therefore judged in relation to others, and often several classification techniques are applied to the same data sets with results compared afterwards.

The most commonly used classification methods in ecological land classification include polythetic hierarchical agglomerative clustering and polythetic hierarchical divisive clustering. The term polythetic means that information on all variables is used to assign observations to a cluster, as opposed to earlier monothetic methods that used single variables in a non-multivariate analysis (Gauch 1982).

Polythetic agglomerative clustering has two steps. In ecological studies, the species-by-species data matrix is first used to compute a samples-by-samples dissimilarity matrix. The matrix may be constructed using any one of several distance measures such as Euclidean distance or percent dissimilarity (Jongman et al. 1987). Second, an agglomeration procedure is applied successively to build up a hierarchy of increasingly large clusters, starting with clusters consisting of a single member, and agglomerating these hierarchically until finally a single cluster contains all the species (Gauch 1982).

In agglomerative clustering methods, the groups that closest resemble each other are joined or “linked” first. The definition of closest, or dissimilarity, between groups differs among methods; the linkage method selected determines the dissimilarity, or distance between two clusters (Jongman et al. 1987). The most common linkage methods used in ecological studies include complete-linkage (or furthest-neighbor), single-linkage (or nearest neighbor), average-linkage, and Ward's method (or minimum-variance) clustering (Digby and Kempton 1987). Flexible-beta linkage has also been found to recover underlying cluster structures efficiently (Milligan and Cooper, 1988).

A methodological problem in applied clustering involves the decision of whether or not to standardize the input variables prior to the computation of a Euclidean distance or other dissimilarity measure (Milligan and Cooper, 1988, Stoddard 1979). A number of transformations are commonly used, including logarithmic

transformations, square-root transformations, exponential transformations, range transformations, standardization to mean of zero and unit variance, and transformation to an ordinal scale (Jongman et al. 1987). Cooper (1988) found that transformation by the range of a variable gave consistently superior recovery of the underlying cluster structure.

Standard clustering algorithms can fail to identify clear cluster structure if that structure is confined to a subset of variables (Fowlkes et al. 1988). The inclusion of unnecessary variables in cluster analysis can therefore mask cluster structure, resulting in mixed assignments of observations. Consequently, superfluous and redundant variables are often screened through exploratory data analyses including descriptive statistics and ordination, regression analysis, or other methods prior to clustering. Cluster analysis will also "recover" natural groups even if none exist. If strong, distinct groups do exist then they should be recovered by a number of different linkage methods. If group membership changes with different linkage methods, it indicates weak group structure.

The simplest polythetic divisive classification is subjective ordination space partitioning (Hill 1977). Sample points are positioned in low-dimensional ordination space and partitions are placed subjectively by drawing lines through sparse regions of the cloud of sample points.

Two-way indicator species analysis (TWINSPAN; Hills 1979) is another polythetic divisive clustering technique. TWINSPAN begins with all species or samples (depending on the objectives) in a single cluster, and divides these into smaller clusters by first ordinating data by CA. Species characterizing the CA axis extremes are emphasized to polarize the samples, and the samples are then divided into two clusters by breaking the ordination axis near its middle. This procedure is repeated until each cluster has no more than a chosen minimum number of members. A corresponding species classification is produced, and the samples and species hierarchical classifications are used together to produce an arranged data matrix similar to a Braun-Blauquet table.

Van Groenewoud (1992) asserts that TWINSPAN does not perform a cluster analysis, and thus does not discern spatial vegetative patterns. He concludes that the usefulness and reliability of TWINSPAN depends on how well CA extracts axes that have ecological meaning, how well the CA axes are divided into meaningful segments, and how closely species are associated with certain segments of the multivariate coenoplane.

Literature Cited

Braun-Blanquet, J. 1932. Plant Sociology: the study of plant communities (authorized English translation of Pflanzensoziologie, translated, revised, and edited by G. D. Fuller and H. S. Conrad). McGraw-Hill, New York, 439 pp.

Bray, J.R. and J.T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Mono., 50:131-151.

Causton, D.R. 1988. An introduction to vegetation analysis: principles, practice, and interpretation. Unwin Hyman, London. 342 pp.

Chatfield, C., and A.J. Collins. 1980. Introduction to multivariate analysis. Chapman and Hall, Ltd., London. 246 pp.

Digby, P.G.N., and R.A. Kempton. 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, Ltd. London.

Fowlkes, E.B, R. Gnanadesikan, and J.R. Kettenring. 1988. Variable Selection in Clustering. J.of Classification 5:205-228.

Gauch, H.G. 1982. Multivariate analysis in community ecology. Cambridge University Press, London and New York. 574 p.

Greenacre, M.J. 1993. Correspondence analysis in practice. Academic Press, London.

Greenacre, M.J. 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Hauser, M., and L. Mucina. 1991. Spatial interpolation methods for interpretation of ordination diagrams. Coenoses 6:151-68.

Hill, M.O., and H.G. Gauch. 1980. Detrended correspondence analysis: an improved ordination technique. Vegetatio 42:47-58.

Hill, M.O. 1979. TWINSPAN - a FORTRAN program for arranging multivariate data on an ordered two-way table by classification of the individuals and attributes. Section of Ecology and Systematics, Cornell University, Ithaca, New York.

Host, G.E.; P.L. Polzer, D.J. Mladenoff, M.A. White, T.R. Crow. 1996. A quantitative approach to developing regional ecosystem classifications. Ecological Applications. 6(2):608-618.

Host, G.E., C.W. Ramm, E.A. Padley, K.S. Pregitzer, J.B. Hart, and D.T. Cleland. 1993. Field sampling and data analysis methods for development of ecological land classifications: an application on the Manistee National Forest. USDA Forest Service, North Central Forest Experiment Station General Technical Report NC-162. 47 p.

Host, G.E. and K.S. Pregitzer. 1992. Geomorphic influences on ground-flora and overstory composition in upland forests of northwestern lower Michigan. 1992. Can. J. For. Res. 22:1547-1555.

Host, G.E., K.S. Pregitzer, C.W. Ramm, D.T. Lusch, and D.T. Cleland 1988. Variations in overstory biomass among glacial landforms and ecological land units in northwestern Lower Michigan. Can. J. For. Res. 18:659-668.

Host, G.E., K.S. Pregitzer, C.W. Ramm, J.B. Hart, and D.T. Cleland. 1987. Landform mediated differences in successional pathways among upland forest ecosystems in northwestern Lower Michigan. For. Sci. 33:445-457.

Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren, editors. 1987. Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen, The Netherlands.

Krzanowski, W.J. 1988. Principles of multivariate analysis: a user's perspective. Clarendon Press, Oxford.

Liebart, L., A. Morineau and K. M. Warwick. 1984. Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices. John Wiley and Sons, New York. 231 pp.

Milligan, G.W. and M.C. Cooper. 1988. A Study of Standardization of Variables in Cluster Analysis. Jor. of Classification. 5:181-204.

Morrison, D.F. 1976. Multivariate statistical methods. McGraw-Hill, USA. 415 pp.

Morzuch, B.J. and G.A. Ruark. 1991. Principal components regression to mitigate the effects of multicollinearity. For. Sci. 37:191-199.

Mueller-Dombois, D., Ellenberg, H. 1974. Aims and methods of vegetation ecology. John Wiley and Sons, New York. 547 p.

Newcomer, J.A. 1984. Principal components analysis of total tree form in seven central Pennsylvania hardwood species. Forest Sci. 30: 64-70.

Oksanen, J. 1988. A note on the occasional instability of detrending in correspondence analysis. Vegetatio 82:29-32

Oksanen, J., and P. Huttunen. 1989. Finding a common ordination for several data sets by individual differences scaling. Vegetatio 83:137-45.

Padley, E.A. 1989. Associations among glacial landforms, soils, and vegetation in northeastern lower Michigan. Ph.D. Dissertation, Michigan State University, East Lansing, MI. 202 pp.

Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74:2215-30.

Peet, R.K., R.G. Knox, J.S. Case, and R.B. Allen. 1987. Putting things in order: the advantages of detrended correspondence analysis. Am. Nat. 131 (6):924-934.

Pielou, E.C. 1984. The interpretation of ecological data: A primer on classification and ordination. John Wiley and Sons, New York. 263 pp.

Pregitzer, K.S. and B.V. Barnes. 1982. The use of ground flora to indicate edaphic factors in upland ecosystems of the McCormick Experimental Forest, Upper Michigan. Can. J. For. Res. 12:661-672.

Pregitzer, K.S. and B.V. Barnes. 1984. Classification and comparison of upland hardwood and conifer ecosystems of the Cyrus H. McCormick Experimental Forest, upper Michigan. Can. J. For. Res. 14:362-375.

Retuerto, R., and A. Carballeira. 1991. Defining phytoclimatic units in Galicia, Spain, by means of multivariate methods. J. Veg. Sci. 2:699-710.

Spies, T.A., and B.V. Barnes. 1985a. A multifactor ecological classification of the northern hardwood and conifer ecosystems of Sylvania Recreation Area, Upper Peninsula, Michigan. Can. J. For. Res. 15:949-960.

Stoddard, A.M. 1979. Standardization of Measures Prior to Cluster Analysis. Biometrics 35:765-773.
ter Braak, C. J. F. 1988. CANOCO - an extension of DECORANA to analyze species-environment relationships. Vegetatio 75:159-60.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167-79.

van Groenewoud, H. 1992. The robustness of Correspondence, Detrended Correspondence, and TWINSPAN Analysis. J. Veg Sci. 3:239-246.

Zak, D.R. and K.S. Pregitzer. 1990. Spatial and temporal variability of nitrogen cycling in northern lower Michigan. For. Sci. 36:367-380.

Zak, D.R., K.S. Pregitzer, and G.E. Host. 1986. Landscape variation in nitrogen mineralization and nitrification. Can. J. For. Res. 16:1258-1263.