Multivariate Analysis Techniques in Environmental Science

Mohammad Ali Zare Chahouki
Department of Rehabilitation of Arid and Mountainous Regions, University of Tehran, Iran

1. Introduction

One of the characteristics of environmental data, many of them and the complex relationships between them. To reduce the number variables, different statistical methods exist. Multivariate statistics is used extensively in environmental science. It helps ecologists discover structure and previous relatively objective summary of the primary features of the data for easier comprehension. However, it is complicated in theorical structure and in operational methodology.

In this chapter some important statistical methods such as Principal component analysis (PCA), Canonical correspondence analysis (CCA), Redundancy analysis (RDA), Cluster analysis, and Discriminate function analysis will be explained briefly.

This chapter too cover the statistical analysis of assemblage data (species by samples matrices of abundance, area cover etc) and/or multi variable environmental data which arise in a wide range of applications in ecology and environmental science, from basic ecological studies (e.g. of dietary composition or population size-structure), through community-based field studies, environmental impact assessments and monitoring of large-scale biodiversity change, to purely physical or chemical analyses.

The use of multivariate analysis has been extended much more widely over the past 20 years. Much more is included on techniques such as Canonical Correspondence Analysis (CCA) and Non-metric Multidimensional Scaling (NMS) and another technique to include organisms and organism-environment relationships other than vegetation. Spatially constrained data analysis will be introduced and the importance of accounting for spatial autocorrelation will be emphasized. Use of the methods within ecology and in environmental reconstruction will also be covered. A study and review of the application of multivariate analysis in biogeography and ecology is provided in: Kent, M. (2006).

2. Landscape ecology

Landscape is simply an area of land (at any scale) containing an interesting pattern that affects and is affected by an ecological process of interest. Landscape ecology, then, involves the study of these landscape patterns, the interactions among the elements of this pattern, and how these patterns and interactions change over time. In addition, landscape ecology involves the application of these principles in the formulation and solving of real-world problems (Turner et al, 2001).

Landscape ecology is perhaps best distinguished by its focus on: 1) spatial heterogeneity, 2) broader spatial extents than those traditionally studied in ecology, and 3) the role of humans in creating and affecting landscape patterns and process (Turner et al, 2001).

In effect the role of ecology, and especially that of vegetation science, has been mainly restricted to the evaluation of the landscape with respect to particular demands: either evaluation as an assessment of the qualities of the ecosystem or evaluation as a socio-economic procedure intended to estimate the functions the natural environment fulfills for human societ (Van der Ploeg and Vlijm, 1978).

Landscape ecology theory stresses the role of human impacts on landscape structures and functions. It also proposes ways for restoring degraded landscapes. Landscape ecology explicitly includes humans as entities that cause functional changes on the landscape (Mielke& Berry, 2001).

Landscape ecology theory includes the landscape stability principle, which emphasizes the importance of landscape structural heterogeneity in developing resistance to disturbances, recovery from disturbances, and promoting total system stability (Mielke& Berry, 2001). This principle is a major contribution to general ecological theories which highlight the importance of relationships among the various components of the landscape. Integrity of landscape components helps maintain resistance to external threats, including development and land transformation by human activity. Analysis of land use change has included a strongly geographical approach which has led to the acceptance of the idea of multifunctional properties of landscapes (Mielke et al, 1976). There are still calls for a more unified theory of landscape ecology due to differences in professional opinion among ecologists and its interdisciplinary approach (Bastian 2001).

Landscape ecology is distinguished by its focus on broader spatial extents than those traditionally studied in ecology. This stems from the anthropocentric origins of the discipline.

Despite early attention to the effects of sample area on measurements, such as species-area relationships, the importance of scale was not widely recognized until the 1980's. Recognition that pattern-process relationships vary with scale demanded that ecologist give explicit consideration to scale in designing experiments and interpreting results.

It became evident that different problems require different scales of study, and that most problems require multiple scales of study. The theory of scale and hierarchy emerged as a framework for dealing with scale. The emergence of scale and hierarchy theory provided a partial theoretical framework for understanding pattern-process relationships, which became the basis for the emergence of landscape ecology as a discipline (Turner MG. 1989).

3. Sampling methods

Sampling design varies considerably with habitat type and specific taxonomic groups. Sampling design begins with a clear statement of the question(s) being asked. This may be the most difficult part of the procedure because the quality of the results is dependent on the nature of the original design.

If the sampling is for densities of organisms then at least five replicate samples per sampling site are needed because many statistical tests require that minimal number. Better yet, consider 20 replicates per sampling site and in some cases 50 or more. If sample replicates are less than five then bootstrapping techniques can be used to analyze the data (name). Some type of random sampling should be attempted (e.g., stratified random sampling) or a line intercept method used to estimate densities (e.g., Strong Method).

Measurements of important physical–chemical variables should be made (e.g.,temperature, salinity, sediment grain size, etc...). Field experiments need to be carried out with carefully designed controls. The correct spatial scale needs to be considered when planning experiments (Stiling, 2002).

If vegetation is correlated with geomorphologic landforms, for this reason stratified random sampling method is better and this method employ (Greig-Smith, 1983; Ludwig and Reynold, 1988)

Environmental impact assessments ideally attempt to compare before and after studies. There were differences in sampling sites, sampling dates, effort, replication, taxonomic categories, and recovery data. (Ferson et al., 1986).

Gotelli and Ellison (2004) and Odum and Barrett (2005) studied on sampling design. Diserud and Aagaard (2002) present a method that tests for changes in community structure based on repeated sampling. This may be the plot method of sampling generally consists of three major types:

(1) Simple random or random sampling without replacement, (2) stratified random, and (3) systematic (Cochran, 1977)

Simple random sampling with replacement is inherently less efficient than simple random sampling without replacement (Thompson, 2002). It is important not to have to determine whether any unit in the data is included more than once. Simple random sampling consists of using a grid or a series of coordinate lines (transects) and a table of random numbers to select several plots (quadrats), the size depending on the dimensions and densities of the organisms present.

4. Classification methods

4.1 Measures of similarity and difference (similarity and dissimilarity)

Dissimilarity or distance measures can be categorized as metric, semi-metric, or nonmetric (McCune et al., 2002). Symmetric distances are extremely useful in community ecology. There are different Similarity coefficients based on binary data. Two methods that are simple and give good results are Jaccard and Czekanowski (Sorenson) coefficients range 0 to 1.o.

Following equations are for Jaccard (1901) and Sorenson similarity coefficients

$$J = \frac{c}{a + b + c}$$

Where J= Jcacard coeffiecient, a= number of occurrences of species a alone, b= number of occurrences of b alone, c= number of co-occurences of two species (a and b).

$$C = \frac{2c}{2a + b + c}$$

Where: (same as in Jaccard)

There are different Dissimilarity coefficients based on meristic or metric data. Coefficient that are widely used in ecology studies are Bray-Curtis (1975) dissimilarity coefficient and Morisita’s Index.

Bray-Curtis is recommended by Clarke and Warwick (2001) and others as the most appropriate dissimilarity coefficient for community studies.

$$BC = \frac{\sum_{j=1}^{n} \left| X_{1j} - X_{2j} \right|}{(X_{1j} + X_{2j})}$$

Where
Σ= sum (from 1 to n)
X1j= # organisms of species j (attribute) collected at site 1 (entity).
X2j= # organisms of species j collected at site 2.
BC = Bray – Curtis coefficient of distance
| | = absolute value
J= 1 to n
N= number of species
Krebs (1999) considers Morisita' s index as the best similarity index, as follows:

$$C_{\lambda} = \frac{2\sum X_{ij} X_{ik}}{(\lambda_1 + \lambda_2) N_j N_k}$$

$$\lambda_1 = \frac{\sum \left[X_{ij}(X_{ij} - 1) \right]}{N_i (N_i - 1)}$$

$$\lambda_2 = \frac{\sum [X_{ik}(X_{ik} - 1)]}{N_k (N_k - 1)}$$

Where
Σ= sum
Cm= Morisita' s Index of Similarity
Xij= No. individuals of species I in sample j
Xik= No. individuals of species I in sample k
Nj= Total No. individuals in sample j
Nk= Total No. individuals in sample k
And Coefficients of Association are two types, ranging from -1 to +1 (e.g., Pearson Correlation Coefficient) and from 0 to X (e.g., χ^2) and applicable to binary and continuous data. Pearson coefficient moment correlation uses conjoint absences, the use of which is inappropriate for comparing sites and appropriate for comparing species (Clarke and Warwick, 2001)
Euclidian Distance Euclidean distance is another measure of distance that can be applied to a site by species matrix. It has been widely used in the past because it is compatible with virtually all cluster techniques.

4.2 Cluster analysis

Clustering is a straightforward method to show association data, however, the confidence of the nodes are highly dependent on data quality, and levels of similarity for cluster nodes is dependent on the similarity index used. Krebs (1989) shows that mean linkage is superior to single and complete linkage methods for ecological purposes because the other two are extremes, either producing long or tight, compact clusters respectively. There are, however, no guidelines as to which mean-linkage method is the best (Swan, 1970).

The purpose of two-way clustering (also known as biclustering) is to graphically expose the relationship between cluster analyses and your individual data points. The resulting graph makes it easy to see similarities and differences between rows in the same group, rows in different groups, columns in the same group, and columns in different groups. You can see graphically how groups of rows and columns relate to each other. Two-way clustering refers to doing a cluster analysis on both the rows and columns of your matrix, followed by graphing the two dendrograms simultaneously, adjacent to a representation of your main matrix. Rows and columns of your main matrix are re-ordered to match the order of items in your dendrogram.

Group Linkage Methods

1. Nearest Neighbor
2. Farthest Neighbor
3. Median
4. Group Average
5. Centroid
6. Ward's Method
7. Flexible Beta
8. McQuitty's Method

Ward's is also know as Orloci's and Minimum Variance Method

Table 1. Major types of hierarchical, agglomerative, polythetic clustering strategies

Cluster analysis can be performed using either presence–absence or quantitative data. Each pair of sites is evaluated on the degree of similarity, and then combined sequentially into clusters to form a dendrogram with the branching point representing the measure of similarity.

4.3 TWINSPAN

The TWINSPAN method (from Two Way Indicator Species Analysis, Hill 1979; Hill et al. 1975) is a very popular method among community ecologists and it was partially inspired by the classificatory methods of classical phytosociology (use of indicators for the definition of vegetation types). Two popular agglomerative polythetic techniques are Group Average and Flexible. McCune et al. (2002) recommend Ward’ s method in addition. Gauch (1982) preferred to use divisive polythetic techniques such as TWINSPAN.

This method works with qualitative data only. In order not to lose the information about the species abundances, the concepts of pseudo-species and pseudo-species cut levels were introduced. Each species can be represented by several pseudo-species, depending on its quantity in the sample. A pseudo-species is present if the species quantity exceeds the corresponding cut level.

TWINSPAN is a program for classifying species and samples, producing an ordered two-way table of their occurrence. The process of classification is hierarchical; samples are successively divided into categories, and species are then divided into categories on the basis of the sample classification. TWINSPAN, like DECORANA, has been widely used by ecologists.

4.4 Indicator species analysis

Indicator species analysis is a divisive polythetic method of numerical classification applicable to large sets of qualitative or quantitative data.

This method provides a simple, intuitive solution to the problem of evaluating species associated with groups of sample units Dufrêne and Legendre’s (1997). It combines information on the concentration of species abundance in a particular group and the faithfulness of occurrence of a species in a particular group. This method produces indicator values for each species in each group. These are tested for statistical significance using a Monte Carlo technique. It requires data from two or more sample units. Indicator values (range 0 for no indication to 100 for perfect indication) are presented for each species. The statistical significance of the maximum indicator value recorded for a given species is generated by a Monte Carlo test. There are many types of indicator species ranging from individual species (Dufrêne & Legendre's ,1997).

The identification of characteristic or indicator species is traditional in ecology and biogeography. Field studies describing sites or habitats usually mention one or several species that characterize each habitat. There is clearly a need for the identification of characteristic or indicator species in the fields of monitoring, conservation, and management. Indicator species add ecological meaning to groups of sites discovered by clustering, they can compare typologies derived from data analysis, identify where to stop dividing clusters into subsets, and point out the main levels in a hierarchical classification of sites.

Indicator species differ from species associations in that they are indicative of particular groups of sites. Good indicator species should be found mostly in a single group of a typology and be present at most of the sites belonging to that group (Hill, 1975).

With Classical problem in community ecology and biogeography, species are the best indicators we have for particular environmental conditions. And in long-term environmental follow-up, conservation, ecological management, researchers are looking for bioindicators of habitat types to preserve or rehabilitate.

McGeoch & Chown (1998) found the indicator value method important to conservation ecosystem because it is conceptually straightforward and allows researchers to identify bioindicators for any combination of habitat types or areas of interest, e.g. existing conservation areas, or groups of sites based on the outcome of a classification procedure.

Indicator species are species that, due to their niche preferences, can be used as ecological indicators of community types, habitat conditions, or environmental changes (McGeoch 1998, Carignan and Villard 2002, Niemi and McDonald 2004).

They are usually determined using an analysis of the relationship between the observed species presence–absence or abundance values in a set of sampled sites and a classification of the same sites (Dufrêne and Legendre 1997).

Finally this method may represent the groups of sites in the classification, in different qualitative characteristics of the ecosystem, such as habitat or community types, environmental or succession states, or the levels of controlled experimental designs.

Since indicator species analysis relates two elements, the species and the groups of sites, it can be used for gaining information on either or both. Indeed, indicator species analysis allows the characterization of the qualitative environmental preferences of the target species (for instance, when the groups are habitat types), and identifyes indicators of particular groups of sites, which can be used in further surveys.

The applications of indicator species analysis are many, including conservation, land management, landscape mapping, or design of natural reserves. Indicator species are commonly referred to as 'diagnostic species' in vegetation studies (Chytr et al. 2002).

4.5 Multi-response permutation procedures (MRPP)

Multi-response Permutation Procedure (MRPP) was introduced by Mielke, Berry, and Johnson (1976) as a technique for detecting the difference between a priori classified groups.

It turned out to be an extremely versatile data-analytic framework from which a number of applications fall out, such as the measurement of agreement, multivariate correlation and association coefcients, and the detection of autocorrelation (see Mielke & Berry, 2001 for a complete coverage of applications of the MRPP framework).

MRPP is a non-parametric method for testing the hypothesis of no difference between two or more groups of entities. The groups must be a priori. For example, one could compare species composition between burned and unburned plots to test the hypothesis of no treatment effect.

Discriminant analysis is a parametric method that can be used on the same general class of questions. However, MRPP has the advantage of not requiring assumptions (such as multivariate normality and homogeneity of variances) that are seldom met with ecological community data (Bakus, 2007).

Multiple Response Permutation Procedure (MRPP) provides a test of whether there is a significant difference between two or more groups of sampling units. This difference may be one of location (differences in mean) or one of spread (differences in within-group distance).

Function MRPP operates on a data frame matrix where rows are observations and responses data matrix. The response(s) may be uni or multivariate. The method is mathematically allied with analysis of variance, in that it compares dissimilarities within and among groups. If two groups of sampling units are really different (e.g. in their species composition), then average of the within-group compositional dissimilarities ought to be less than the average of the dissimilarities between two random collection of sampling units drawn from the entire population.

The MRPP method is simply the overall weighted mean of within-group means of the pairwise dissimilarities among sampling units.

The MRPP algorithm first calculates all pairwise distances in the entire dataset, then calculates delta. It then permutes the sampling units and their associated pairwise distances, and recalculates a delta based on the permuted data. It repeats the permutation step permutations times.

The function also calculates the change-corrected within-group agreement. And it also calculates classification strength which is defined as the difference between average between group dissimilarities and within group dissimilarities (Van Sickle 1997).

If the first argument data can be interpreted as dissimilarities, they will be used directly. In other cases the function treats datas observations, and uses vegdist to find the dissimilarities. The default distance is Euclidean as in the traditional use of the method, but other dissimilarities in vegdist also are available.

Function meandist calculates a matrix of mean within-cluster dissimilarities (diagonal) and between-cluster dissimilarities (off-diagonal elements), and an attribute n of grouping counts. Function summary finds the within-class, between-class and overall means of these dissimilarities, and the MRPP statistics with all weight type options and the classification strength.

The MRPP a robust alternative to the traditional normal theory based parametric tests, such as the t test and the analysis of variance. However, MRPP is not widely known to researchers, and part of the reason is that it has not been incorporated into major statistical packages.

4.6 Mantel test

The Mantel test evaluates the null hypothesis of no relationship between two dissimilarity (distance) or similarity matrices. The Mantel test is an alternative to regressing distance matrices that circumvents the problem of partial dependence in these matrices.

For example evaluating the correspondence between two groups of organisms from the same set of sample units or comparing community structure before and after a disturbance is evaluate by mantel test.

In Consequency, Two methods are available in PC-ORD: Mantel's asymptotic approximation and a randomization (Monte Carlo) method.

The Mantel Test (Mantel, 1967) compares two dissimilarity (distance) matrices using Pearson correlation. The matrices that use in this test must be of the same size (i.e., same number of rows). It would appear that the Mantel Test could be used to evaluate the internal structure in two sets of samples by comparing the two (dis)similarity matrices.

McCune and Mefford (1999) have a computer program for performing the Mantel test. After the Mantel statistic has been calculated, the statistical significance of the relationship is tested by a permutation test or by using an asymptotic t-approximation test.

The minimum number of permutations (randomizations) recommended by Manly (1997) is 1000. The Mantel and ANOSIM procedures produce similar probabilities (Legendre and Legendre, 1998).

Some of the problems associated with the Mantel test include (1) weakness in detecting spatial autocorrelation where the spatial pattern is complex and not easily modeled with distance matrices, (2) a larger number of data points may be needed for field experiments than is usually obtained, and (3) multivariate data are summarized into a single distance or dissimilarity and it is not possible to identify which variable(s) contributed the most (Fortin and Gurevitch, 2001).

The use of Partial Mantel tests can distinguish the relative contributions of the factors of a third matrix considered as covariables.

Mantel and partial Mantel tests can produce complementary information that other methods such ANOVA cannot provide and may do a better job than other method, in detecting block effects (Fortin and Gurevitch, 2001).

Thus, it is possible to distinguish the effects of spatial pattern from those of experimentally imposed treatment effects with these Mantel tests (Bakus, 2007).

4.7 The best clustering strategies

The similarity or dissimilarity measures, forming the backbone of most multivariate clustering and ordination techniques, which are favored by many ecologists, are Jaccard and Bray–Curtis.

The most popular clustering techniques in ecological studies are Group Average and Flexible group average clustering. For many of these techniques, it is impossible to choose a "best method" because of the heuristic nature of the methods (Jongman et al., 1995; Anderson, 2001).

The choice of an index must be made based on the investigator's experience, the type of data collected, and the ecological question to be answered. When comparisons are made, the Jaccard index is among the least sensitive of the similarity (or dissimilarity) indices (Jongman et al. 1995). Similarly, hierarchical agglomerative techniques (e.g., particularly Group Average) have proven to be very useful to ecologists in the construction of dendrograms.

As part of a study is needed information on plant communities and their distribution. Then the objective of best method of multivariate analysis is to:

1. Identify and describe plant communities on the ecosystem
2. Map these communities to provide a tool for ecological studies and monitoring of ecological community;

3. Characterize the interrelationships between plant communities, soil particle size, moisture availability, grazing pressure and elevation.

5. Ordination methods

Ordination is a collective term for multivariate techniques which adapt a multi-dimensional swarm of data points in such a way that when it is projected onto a two dimensional space any intrinsic pattern the data may possess becomes apparent upon visual inspection (Pielou, 1984). As mentioned in the introduction, ordination is the arrangement of samples along gradients. Indeed, ordination can be considered a synonym for multivariate gradient analysis.

Basically, ordination serves to summarize community data by producing a low-dimensional ordination space in which similar species and samples are plotted close together, and dissimilar species and samples are placed far apart.

Indirect gradient analysis
    Distance-based approaches
    Polar ordination, PO (Bray-Curtis ordination)
    Principal Coordinates Analysis, PCoA (Metric multidimensional scaling)
    Nonmetric Multidimensional Scaling, NMDS
Eigenanalysis-based approaches
    Linear model
    Principal Components Analysis, PCA
    Unimodal model
    Correspondence Analysis, CA (Reciprocal Averaging)
    Detrended Correspondence Analysis, DCA
Direct gradient analysis
    Linear model
    Redundancy Analysis, RDA
    Unimodal model
    Canonical Correspondence Analysis, CCA
    Detrended Canonical Correspondence Analysis, DCCA

5.1 Polar ordination (Bray- Curtis)

(Polar Ordination) arranges samples with respect to “poles” (also termed end points or reference points) according to a distance matrix (Bray and Curtis 1957).. These endpoints are two samples with the highest ecological distance between them (objective approach), OR two samples suspected of being at opposite ends of an important gradient (subjective approach). This procedure is especially useful for investigating ecological change (e.g., succession, recovery).

Advantages of this method is Ideal for evaluating problems with discrete endpoints.

Polar Ordination ideal for testing specific hypotheses (e.g., reference condition or experimental design) by subjectively selecting the end points Disadvantages:

This technique does not provide a general-purpose description of the community (perspective is biased) Very sensitive to outliers (by definition – “end points”) Select a distance measure (usually Sorensen Index) and calculate matrix of distances (D) between all pairs of points (Beals, 1984).

In the earliest versions of PO, these endpoints were the two samples with the highest ecological distance between them, or two samples which are suspected of being at opposite ends of an important gradient (thus introducing a degree of subjectivity).

Beals (1984) extended Bray-Curtis ordination and discussed its variants, and is thus a useful reference. The polar ordination, simplest method is to choose the pair of samples, not including the previous endpoints, with the maximum distance of separation.

5.2 Principal component aanalysis

PCA was invented in 1901 by Karl Pearson (Dunn, et al, 1987) Now it is mostly used as a tool in exploratory data analysis and for making predictive models.

Principal Components Analysis is a method that reduces data dimensionality by performing a covariance analysis between factors.

It can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute.

The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular case in the data) and loadings (the weight by which each standarized original variable should be multiplied to get the component score (Feoli and Orl¢ci. 1992).

Principal components analysis is the basic eigenanalysis technique. It maximizes the variance explained by each successive axis.

PCA was one of the earliest ordination techniques applied to ecological data. PCA uses a rigid rotation to derive orthogonal axes, which maximize the variance in the data set.

Both species and sample ordinations result from a single analysis. Computationally, PCA is basically an eigenanalysis. The sum of the eigenvalues will equal the sum of the variance of all variables in the data set. PCA is relatively objective and provides a reasonable but crude indication of relationships.

This method is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components.

The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components.

Principal components are guaranteed to be independent only if the data set is jointly normally distributed. It is sensitive to the relative scaling of the original variables.

Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).

Broken-stick eigenvalues are provided to help evaluating statistical significance. Principal component analysis (PCA) (ter Braak and Sˇ milauer, 1998) was used to determine the association between plant communities and environmental variables, i.e. in an indirect non-canonical way (ter Braak and Loomans, 1987).

While PCA finds the mathematically optimal method (as in minimizing the squared error), it is sensitive to outliers in the data that produce large errors PCA tries to avoid. It therefore is common practice to remove outliers before computing PCA.

However, in some contexts, outliers can be difficult to identify. For example in data mining algorithms like correlation clustering, the assignment of points to clusters and outliers is not known beforehand.

A recently proposed generalization of PCA based on Weighted PCA increases robustness by assigning different weights to data objects based on their estimated relevancy.

Although it has severe faults with many community data sets, it is probably the best technique to use when a data set approximates multivariate normality. PCA is usually a poor method for community data, but it is the best method for many other kinds of multivariate (Bakus, 2007).

5.3 Principal coordinate analysis (PCoA)

Principal Coordinate Analysis (PCoA) is a method to represent on a 2 or 3 dimensional chart objects described by a square matrix containing that contains resemblance indices between these objects.

This method is due to Gower (1966). It is sometimes called metric MDS (MDS: Mutidimensional scaling) as opposed to the MDS (or non-metric MDS). Both methods have the same objective and produce similar results if the similarity matrix f square distances are metric and if the dimensionality is sufficient.

Principle coordinates are similar to principal components in concept. The advantage of PCoA is that it may be used with all types of variables (Legendre and Legendre, 1998). Because of this, PCoA is an ordination method of considerable interest to ecologists.

Most metric (PCoA) and nonmetric MDS plots are very similar or even identical, provided that a similar distance measure is used. The occurrence of negative eigenvalues, lack of emphasis on distance preservation, and other problems are discussed in detail by Legendre and Legendre (1998) and Clarke and Warwick (2001).

One of the biggest differences between PCA and PCoA is that the variables (i.e. species) representing the original axes are projected as biplot arrows. In the bryophyte communities, these biplot arrows greatly aid in interpretation (Bakus, 2007).

In most applications of PCA (e.g. as a factor analysis technique), variables are often measured in different units. For example, PCA of taxonomic data may include measures of size, shape, color, age, numbers, and chemical concentrations. For such data, the data must be standardized to zero mean and unit variance (the typical default for most computer programs).

For ordination of ecological communities, however, all species are measured in the same units, and data should not be standardized. In matrix algebra terms, most PCAs are eigenanalyses of the correlation matrix, but for ordination they should be PCAs of the covariance matrix.

In contrast to Correspondence Analysis and related methods, species are represented by arrows. This implies that the abundance of the species is continuously increasing in the direction of the arrow, and decreasing in the opposite direction. Thus PCA is a ‘linear method’.

Although the discussion above implies that PCA is distinctly different from PCoA, the two techniques end up being identical, if the distance metric is Euclidean.

Unfortunately, this linear assumption causes PCA to suffer from a serious problem, the horseshoe effect, which makes it unsuitable for most ecological data sets (Gauch 1982).

Principal Coordinates Analysis (PCoA, = Multidimensional scaling, MDS) is a method to explore and to visualize imilarities or dissimilarities of data. It starts with a similarity matrix or dissimilarity matrix (= distance matrix) and assigns for each item a location in a low-dimensional space, e.g. as a 3D graphics.

PCOA tries to find the main axes through a matrix. It is a kind of eigenanalysis (sometimes referred as "singular value decomposition") and calculates a series of eigenvalues and eigenvectors. Each eigenvalue has an eigenvector, and there are as many eigenvectors and eigenvalues as there are rows in the initial matrix.

Eigenvalues are usually ranked from the greatest to the least. The first eigenvalue is often called the "dominant" or leading" eigenvalue. Using the eigenvectors we can visualize the main axes through the initial distance matrix. Eigenvalues are also often called "latent values".

The result is a rotation of the data matrix: it does not change the positions of points relative to each other but it just changes the coordinate systems!

By using PCoA we can visualize individual and/or group differences. Individual differences can be used to show outliers.

There is also a method called 'Principal Component Analysis' (PCA, sometimes also misleadingly abbreviated as 'PCoA') which is different from PCOA.

4. Principal Coordinates Analysis (Principal Coordinates Analysis (PCoA PCoA or PCO) or PCO)
5. Maximizes the linear correlation between distance measures and distance in the ordination.
6. useful if one has only a distance (or similarity) matrix
7. The underlying model is that there a fixed number of explanatoryoriginal variables. In contrast, PCA, RA, and DCA assume that there are potentially many variables, but of declining importance.
8. One cannot easily put new points in a PCoA.
9. For Euclidean distance, PCoA= PCA
10. PCoA can be expressed as an eigenanalysis.

5.4 Factor analysis (FA)

FA and PCA (principal components analysis) are methods of data reduction. Take many variables and explain them with a few “factors” or “components”.

Correlated variables are grouped together and separated from other variables with low or no correlation. Patterns of correlations are identified and either used as descriptives (PCA) or as indicative of underlying theory (FA). It process of providing an operational definition for latent construct (through regression equation).

FA and PCA are not much different than canonical correlation in terms of generating canonical variates from linear combinations of variables. Although there are now no “sides” of the equation.

For calcuting this method we should:

1. Selecting and Measuring a set of variables in a given domain
2. Data screening in order to prepare the correlation matrix
3. Factor Extraction
4. Factor Rotation to increase interpretability
5. Interpretation

Factor analysis is seldom used in ecology. Several statisticians state that it should not be used because it is based on a special statistical model.

Estimating the unique variance is the most difficult and ambiguous task in FA (McGarigal et al., 2000).

5.5 Redundancy analysis

RDA is a linear method and since it is a linear method, species as well as environmental variables are represented by arrows. In most cases, it is best to represent the two sets of arrows in two figures for ease of display. Thus, if have a gradient along which all species are positively correlated, RDA will detect such a gradient while CCA will not. RDA can use 'species' that are measured in different units. If so, the data must be centered and standardized. But in fact, as an ordination technique, the species should not be standardized.

Redundancy analysis is the linear method of direct ordination. It also goes under the name of principal components analysis with respect to instrumental variables, which in our case are environmental variables (Sabatier et al., 1989). It is also called least-squares reduced rank regression so as to emphasize its link with multivariate regression (ter Braak & Prentice, 1988; ter Braak & Looman, 1994). In statistical textbooks on multivariate analysis, redundancy analysis is usually neglected.

RDA is useful when gradients are short. In particular, RDA may be the method of choice in a short-term experimental study. In such cases, the treatments are the explanatory variables (and are usually dummy variables). The sample ID or block might be a covariable in a partial RDA, if one wishes to factor out local effects (Bakus, 2007).

'variance explained' is actually a variance explained, and not merely inertia. Thus, variance partitioning, and interpretation of eigenvalues, are more straightforward than for CCA (Lepš & Šmilauer, 1999).

The relatively short space devoted to RDA should not be given as an indication that it is less valuable than, or inferior to, other method. It is simply used for different purposes.

Consequently, the linear method of direct ordination, redundancy analysis, can often efficiently display the interesting effects, and there is no need for methods that also work when the range of community variation is larger (such as canonical correspondence analysis) nor for unconstrained nonmetric multidimensional scaling.

5.6 Correspondence analysis (CA) or reciprocal averaging (RA)

Reciprocal averaging is also known as correspondence analysis (CA) because one algorithm for finding the solution involves the repeated averaging of sample scores and species scores (citations). It is a graphical display ordination technique which simultaneously displays the rows (sites) and columns (species) of a data matrix in low dimensional space (Gittins, 1985). Row identifiers (species) plotted close together are similar in their relative profiles, and column identifiers plotted close together are correlated, enabling one to interpret not only which of the taxa are clustered, but also why they are clustered (Bakus,2007).

Reciprocal averaging (RA) yields both normal and transpose ordinations automatically. Like DCA, RA ordinates both species and samples simultaneously. Instead of maximizing ‘variance explained’, CA maximizes the correspondence between species scores and sample scores.

If species scores are standardized to zero mean and unit variance, the eigenvalues also represent the variance in the sample scores (but not, as is often misunderstood, the variance in species abundance).

Since CA is a unimodal model, species are represented by a point rather than an arrow. This is (under some choices of scaling; see ter Braak and Šmilauer 1998) the weighted average of the samples in which that species occurs. With some simplifying assumptions (ter Braak and Looman 1986), the species score can be considered an estimate of the location of the peak of the species response curve.

The CA distortion is called the arch effect, which is not as serious as the horseshoe effect of PCA because the ends of the gradients are not incurved. Nevertheless, the distortion is prominent enough to seriously impair ecological interpretation (Bakus, 2007).

In other words, the spacing of samples along an axis may not affect true differences in species composition.

Gradient compression can be quite blatant in simulated data sets. The problems of gradient compression and the arch effect led to the development of Detrended Correspondence Analysis (Bakus, 2007).

5.7 Detrended correspondence analysis (DCA)

Detrended Correspondence Analysis (DCA) eliminates the arch effect by detrending (Hill and Gauch 1982). It's a series of rules that are used to reshape data to make it friendlier for analysis. Once again, primarily used for ecological data, but can be extended to anything (data simply can't contain negative values).

The reason that this technique is used is to over come the arch effect (the horseshoe effect).

Found in data whenever "PCA or other distance conserving ordination techniques are applied to data which follow a continuous gradient, along which there is a progressive turnover of dominant variables." Such as in ecological succession

After ordination by a distance conserving technique and the first two axes are plotted against each other, one would find an arch shape.

DCA is another eigenanalysis ordination technique that based on reciprocal averaging (RA; Hill 1979). DCA is geared to ecological sets, is based on samples and species. DCA ordinates both species and samples simultaneously.

There are two basic approaches to detrending: by polynomials and by segments (ter Braak and Šmilauer 1998). Detrending by polynomials is the more elegant of the two: a regression is performed in which the second axis is a polynomial function of the first axis, after which the second axis is replaced by the residuals from this regression. Similar procedures are followed for the third and higher axes.

The compression of the ends of the gradients is corrected by nonlinear rescaling. Rescaling shifts sample scores along each axis such that the average width is equal to 1.

Rescaling has a beneficial consequence: the axes are scaled in units of beta diversity (SD units, or units of species standard deviations). Thus if the underlying gradient is important well known, it is possible to plot the DCA scores as a function of the gradient, and there by determine whether the species 'perceive' the gradient differently than we measure it.

The shape of the species response curves may change if axes are rescaled. Thus, skewness and kurtosis are largely artifacts of the units of measurement for which we choose to measure the environment.

5.8 Nonmetric multimentional scaling (MDS, NMDS, NMS, NMMDS)

Nonmetric Multidimensional Scaling (NMDS) rectifies this by maximizing the rank order correlation. For this proceeds at first the user selects the number of dimensions (N) for the solution, and chooses an appropriate distance metric and then The distance matrix is calculated. And initial configuration of samples in N dimensions is selected. This configuration can be random, though the chances of reaching the correct solution are enhanced if the configuration is derived from another ordination method.

And finally, the final configuration of points represents your ordination solution. The configuration is dependent on the number of dimensions selected; e.g. the first two axes of a 3-dimensional solution does not necessarily resemble a 2-dimensional solution. The stress will typically decrease as a function of the number of dimensions chosen; this function can aid in the selection of the results (Bakus, 2007).

This is why it is sometimes useful to rotate the solution (such as by the Varimax method) – although there is no theory that states that the final solution will represent a 'gradient' Other problems and advantages of NMDS will be discussed later, when comparing it to Detrended Correspondence Analysis (Bakus, 2007).

5.9 MANOVA and MANCOVA

A factorial MANOVA may be used to determine whether or not two or more categorical grouping variables (and their interactions) significantly affect optimally weighted linear combinations of two or more normally distributed outcome variables.

These parametric multivariate techniques (Multivariate Analysis of Variance and Multivariate Analysis of Covariance) are similar to ANOVA and ANCOVA MANOVA (Wilks' Lambda) and ANCOVA are advantageous in that performing multiple univariate tests can inflate the a value, leading to false conclusions (Scheiner, 2001).

MANOVA seeks differences in the dependent variables among the groups (McCune et al, 2002). Assumptions of MANOVA include multivariate normality (error effects included), independent observations, and equality of variance-covariance matrices (Paukert and Wittig, 2002). Because of these assumptions, among others, MANOVA is not often used in ecology although its use in increasing.

The power of traditional MANOVA declines with an increase in the number of response variables (Scheiner, 2001). Unequal sample sizes are not a large problem for MANOVA, but may bias the results for factorial or nested designs. Before ANCOVA is run, tests of the assumption of homogeneity of slopes need to be performed (Petratis et al., 2001).

Early attempts to develop nonparametric multivariate analysis include those of Mantel and Valand (1970). They are more complex as they handle three or more variables simultaneously. They are frequently used with the analysis of experimental studies, especially laboratory experiments.

More recently, the Analysis of Similarities was developed to compare communities or changes in communities because of pollution (Clarke, 1993). For typical species abundance matrices, an Analysis of Similarities (ANOSIM) permutation procedure is recommended over MANOVA.

Multiple analysis of variance (MANOVA) is used to see the main and interaction effects of categorical variables on multiple dependent interval variables. MANOVA uses one or more categorical independents as predictors, like ANOVA, but unlike ANOVA, there is more than one dependent variable. Where ANOVA tests the differences in means of the interval dependent for various categories of the independent(s), MANOVA tests the differences in the centroid (vector) of means of the multiple interval dependents, for various categories of the independent(s). One may also perform planned comparison or post hoc comparisons to see which values of a factor contribute most to the explanation of the dependents.

There are multiple potential purposes for MANOVA.

To compare groups formed by categorical independent variables on group differences in a set of interval dependent variables.

To use lack of difference for a set of dependent variables as a criterion for reducing a set of independent variables to a smaller, more easily modeled number of variables.

Multiple analysis of covariance (MANCOVA) is similar to MANOVA, but interval independents may be added as "covariates." These covariates serve as control variables for the independent factors, serving to reduce the error term in the model. Like other control procedures, MANCOVA can be seen as a form of "what if" analysis, asking what would happen if all cases scored equally on the covariates, so that the effect of the factors over and beyond the covariates can be isolated. The discussion of concepts in the ANOVA section also applies, including the discussion of assumptions.

5.10 Discriminate analysis

Discriminate Analysis (DA) is a powerful tool that can be used with both clusters of species data and environmental variables. It determines which variables discriminate between two or more groups, that is, independent variables are used as predictors of group membership (McCune et al., 2002). It is very similar to MANOVA and multiple regression analysis (Statsoft, Inc., 1995; McGarigal et al., 2000). Clusters can be identified by several methods using raw data: (1) constructing a dendrogram, (2) using PCA (even if you have field data) for initial visual identification of clusters, and (3) point rotation in space by rotating ordinations (i.e., rotating axes – see McCune and Mefford, 1999). If any method indicates groups or clusters of data then DA can be used. However, the number of groups is set before the DA analysis. DA finds a transform for the minimum ratio of difference between pairs of multivariate means and variances in which two clusters are separated the most and inflated the least.

DA produces two functions: (1) classification function consisting of 2 groups or clusters of points (this information can be used for prediction with probabilities) and (2) discriminate function containing environmental variables that can be used to discriminate differences among the groups.

DA differs from PCA and Factor Analysis in that no standardization of data is needed (PCA and FA need standardization because of scaling problems) and the position of the axes distinguishes the maximum distance between clusters. (Davis, 1986).

DA assumes a multivariate normal distribution, homogeneity of variances, and independent samples (Paukert and Wittig, 2002). Violations of normality are usually not fatal (i.e., somewhat non-normal data can be used). A description of the procedure to use DA with Statistica is given in the Appendix. Multiple Discriminate Analysis (MDA) is the term often used when three or more clusters of data are processed simultaneously. MDA is particularly susceptible to rounding error. Calculations in double precision for at least the eigenvalue-eigenvector routines are advisable (Green, 1979). Limitations of Discriminate Analysis are discussed by McGarigal et al. (2000) Dytham (1999).

5.11 Canonical correspondence analysis (CCA)

In ecology studies, the ordination of samples and species is constrained by their relationships to environmental variables. When species responses are unimodal (hump-shaped), and by measuring the important underlying environmental variables, CCA is most likely to be useful.

It was used to examine the relationships between the measured variables and the distribution of plant communities (Ter Braak, 1986). CCA expresses species relationships as linear combinations of environmental variables and combines the features of CA with canonical correlation analysis (Green, 1989). This provides a graphical representation of the relationships between species and environmental factors.

Canonical Correlation Analysis is presented as the standard method to relate two sets of variables (Gittins, 1985). However, the latter method is useless if there are many species compared to sites, as in many ecological studies, because its ordination axes are very unstable in such cases.

The best weight for CCA describes environment variables with the first axis shows. Species information structure using a reply CCA Nonlinear with the linear combination of variables will consider environmental characteristics of acceptable behavior characteristics of species with environment shows. CCA analysis combined with non-linear species and environmental factors shows the most important environmental variable in connection with the axes shows.

In Canonical Correspondence Analysis, the sample scores are constrained to be linear combinations of explanatory variables. CCA focuses more on species composition, i.e. relative abundance.

When a combination of environmental variables is highly related to species composition, this method, will create an axis from these variables that makes the species response curves most distinct. The second and higher axes will also maximize the dispersion of species, subject to the constraints that these higher axes are linear combinations of the explanatory variables, and that they are orthogonal to all previous axis.

Monte Carlo permutation tests were subsequently used within canonical correspondence analysis (CCA) to determine the significance of relations between species composition and environmental variables (ter Braak, 1987)

The outcome of CCA is highly dependent on the scaling of the explanatory variables. Unfortunately, we cannot know a priori what the best transformation of the data will be, and it would be arrogant to assume that our measurement scale is the same scale used by plants and animals. Nevertheless, we must make intelligent guesses (Bakus, 2007).

In CCA possible that patterns result from the combination of several explanatory variables. And many extensions of multiple regression (e.g. stepwise analysis and partial analysis) also apply to CCA.

It is possible to test hypotheses (though in CCA, hypothesis testing is based on randomization procedures rather than distributional assumptions). Explanatory variables can be of many types (e.g. continuous, ratio scale, nominal) and do not need to meet distributional assumptions.

Another advantage of CCA lies in the intuitive nature of its ordination diagram, or triplot. It is called a triplot because it simultaneously displays three pieces of information: samples as points, species as points, and environmental variables as arrows (or points).

If data sets are few, CCA triplots can get very crowded then should be separate the parts of the triplot into biplots or scatterplots (e.g. plotting the arrows in a different panel of the same figure) or rescaling the arrows so that the species and sample scores are more spread out. And we can only plotting the most abundant species (but by all means, keep the rare species in the analysis).

Noise in the species abundance data set is not much of a problem for CCA (Palmer, 1988). However, it has been argued that noise in the environmental data can be a problem (McCune 1999). It is not at all surprising that noise in the predictor variables will cause noise in the sample scores, since the latter are linear combinations of the former.

It is probably obvious that the choice of variables in CCA is crucial for the output. Meaningless variables will produce meaningless results. However, a meaningful variable that is not necessarily related to the most important gradient may still yield meaningful results (Palmer, 1988).

If many variables are included in an analysis, much of the inertia becomes ‘explained’. Any linear transformation of variables (e.g. kilograms to grams, meters to inches, Fahrenheit to Centigrade) will not affect the outcome of CCA whatsoever.

There are as many constrained axes as there are explanatory variables. The total ‘explained inertia’ is the sum of the eigenvalues of the constrained axes. The remaining axes are unconstrained, and can be considered ‘residual’. The total inertia in the species data is the sum of eigenvalues of the constrained and the unconstrained axes, and is equivalent to the sum of eigenvalues, or total inertia, of CA. Thus, explained inertia, compared to total inertia, can be used as a measure of how well species composition is explained by the variables. Unfortunately, a strict measure of ‘goodness of fit’ for CCA is elusive, because the arch effect itself has some inertia associated with it (Bakus, 2007).

The ordination diagrams of canonical correlation analysis and redundancy analysis display the same data tables; the difference lies in the precise weighing of the species (ter Braak & Looman, 1994; Van der Myer, 1991)

One of limitations to CCA is that correlation does not imply causation, and a variable that appears to be strong may merely be related to an unmeasured but ‘true’ gradient. As with any technique, results should be interpreted in light of these limitations (McCune, 1999).

5.12 Multiple regression (MR) (multiple linear regression)

The mechanics of testing the "significance" of a multiple regression model is basically the same as testing the significance of a simple regression model, we will consider an F-test, a t-test (multiple t's) and R-sqrd. However, unlike simple regression where the F & t tests tested the same hypothesis, in multiple regression these two tests have different purposes. R-sqrd is still the percent of variance explained but is no longer the correlation squared (as it was with in simple linear regression) and we will also introduce adjusted R-sqrd. When considering a multiple regression (MR) model the most common order to interpret things consists of first looking at the R-sqrd, then testing the entire model by looking at the F-test, and finally looking at each individual coefficient individually using the t-tests. NOTE: The term "significance" is a nice convenience but is very ambiguous in definition if not properly specified. Thus when taking this class you should avoid simply saying something is significant without explaining (1) how you made that determination, and (2) what that specifically means in this case. You will see from the examples that those two things are always done. If you cannot do that then any time you use the word "significant" you are potentially hurting yourself in two ways; (1) you won't do well on the quizzes or exams where you have to be able to be more explicit than simply throwing out the word "significant", and (2) you will look like a fool in the business world when somebody asks you to explain what you mean by "significant" and you are stumped. Remember if you can't explain your results in managerial terms than you do not really understand what you are doing.

In order to show the relationship between biotic (principal component axes) and abiotic factors (environmental factors), a multiple regression type analysis is used. Multiple regression solves simultaneously normal equations and produces partial regression coefficients. Partial regression coefficients each give the rate of change or slope in the dependent variable for a unit of change in a particular independent variable, assuming all other independent variables are held constant. MR is not considered by some statisticians as a multivariate procedure because it includes only one dependent variable (Paukert and Wittig, 2002).

The objective of multiple regression is to determine the influence of independent variables on a dependent variable, for example, the effect of depth, sediment grain size, salinity, temperature, and predator density on the population density of species.

The parameters are estimated by the least-squares method, that is, minimizing the sum of squares of the differences between the observed and expected response (Jongman et al., 1995)

Normally component loadings (e.g., scores in PCA) suggest which variables are most important. However, with species abundances the only variable in some ordinations, one must use other techniques to attempt to suggest what may have produced the gradients for Axis 1, 2, 3, and so forth.

Univariate analyses such as the Spearman rank correlation coefficient are not as ecologically realistic as multivariate analyses such as multiple regression because some variables are correlated and there are interaction effects between variables (Jongman et al., 1995).

The highest standardized partial regression coefficients (positive or negative) suggest the most important factors (e.g., sediment size, predator density, etc.) in controlling the population density of species X. Significance tests and standard errors then can be calculated from the data.

Multiple regression has many potential problems such as the type of response curve, error distribution, and outliers that may unduly influence the results (Jongman et al., 1995). Multiple regression variables may be highly correlated, therefore, examine the correlation coefficients first (i.e., run a multiple correlation analysis between variables) to exclude some of them before doing multiple regressions.

Multiple regression generally should employ a maximum of 6 variables. Legendre and Legendre (1998) suggest a stepwise procedure for reducing numerous variables. This involves a process of alternating between forward selection and backward elimination (Kutner et al., 1996).

Lee and Sampson (2000) took ordination scores, representing a gradient in fish communities, and regressed them against a group of environmental variables and time. Many of scientist study on multiple regression and computer program (Such as Davis, 1986; and Sokal and Rohlf, 1995)

In multiple regression, it is typical to include quadratic terms for explanatory variables. For example, if you expect a response variable to reach a maximum at an intermediate value of an explanatory variable, including this explanatory variable AND the square of the explanatory variable may allow a concave-down parabola to provide a reasonable fit.

This is an analogous situation to multiple regression: the multiple r^2 or 'variance explained' increases as a function of the number of variables included.

Both multiple regression and CCA find the best linear combination of explanatory variables, they are not guaranteed to find the true underlying gradient (which may be related to unmeasured or unmeasurable factors), nor are they guaranteed to explain a large portion of variation in the data.

5.13 Path analysis

In statistics, path analysis is used to describe the directed dependencies among a set of variables. This includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminate analysis, as well as more general families of models in the multivariate analysis of variance and covariance analyses (MANOVA, ANOVA, ANCOVA).

Path analysis is a straightforward extension of multiple regression. Its aim is to provide estimates of the magnitude and significance of hypothesised causal connections between sets of variables. This is best explained by considering a path diagram.

Path analysis is an extension of multiple linear regression, allowing interpretation of linear relationships among a small number of descriptors (Legendre and Legendre, 1998).

This method was originally developed by Sewall Wright in which he introduced the concept of a path diagram. It handles more than one dependent variable and the effects of dependent variables on one another (Mitchell, 2001).

Path analysis assumes a normal distribution of residuals, additive and linear effects, inclusion of all important variables, that residual errors are uncorrelated, and that there is no measurement error.

Path analysis was developed as a method of decomposing correlations into different pieces for interpretation of effects (e.g., how does parental education influence children's income 40 years later?). Path analysis is closely related to multiple regression; you might say that regression is a special case of path analysis.

Some people call this stuff (path analysis and related techniques) causal modeling. The reason for this name is that the techniques allow us to test theoretical propositions about cause and effect without manipulating variables. However, the causal in causal modeling refers to an assumption of the model rather than a property of the output or consequence of the technique. That is, people assume some variables are causally related, and test propositions about them using the techniques.

5.14 Canonical correlation analysis (CVA)

Canonical variate analysis (CVA) is a widely used method for analyzing group structure in multivariate data. It is mathematically equivalent to a one-way multivariate analysis of variance and often goes by the name of canonical discriminate analysis. Change over time is a central feature of many phenomena of interest to researchers. This dissertation extends CVA to longitudinal data. It develops models whose purpose is to determine what is changing and what is not changing in the group structure. Three approaches are taken: a maximum likelihood approach, a least squares approach, and a covariance structure analysis approach. All methods have in common that they hypothesize canonical variates which are stable over time.

The maximum likelihood approach models the positions of the group means in the subspace of the canonical variates. It also requires modeling the structure of the within-groups covariance matrix, which is assumed to be constant or proportional over time.

In addition to hypothesizing stable variates over time, one can also hypothesize canonical variates that change over time. Hypothesis tests and confidence intervals are developed. The least squares methods are exploratory. They are based on three-mode PCA methods such as the Tucker2 and parallel factor analysis. Graphical methods are developed to display the relationships between the variables over time.

Stable variates over time imply a particular structure for the between-groups covariance matrix. This structure is modeled using covariance structure analysis, which is available in the SAS package Proc Calis.

Canonical Variate Analysis is a special case of Canonical Correlation Analysis (Jongman et al., 1995). It is also described as a type of linear discriminate analysis (McGarigal et al., 2000). The set of environmental variables consists of a single nominal variable defining the classes.

CVA is usable only if the number of sites is much greater than the number of species and the number of classes. Many ecological data sets cannot be analyzed by CVA without dropping many species, thus CVA is not used much in ecology.

Instead, they usually give the impression that there is only one such hypothesis, and therefore only one statistical technique is needed-Hotelling's canonical variate analysis (CVA).

Most discussions of CVA are restricted almost entirely to a description of the underlying mathematical theory, computing directions, and perhaps an example of the computations. Very little is usually said about the logic of the method, so that the reader is unable to judge for himself whether the method described can actually be used to test the hypothesis of interest to him.

Even when CVA is appropriate, several prominent sources have recommended misleading interpretations of the statistics computed in CVA.

Perhaps because of these deficiencies, many behavioral scientists have concluded incorrectly that CVA has few or no valid and important uses in the behavioral sciences. The use originally proposed by Hotelling has been rejected by most behavioral scientists.

6. Ordination and classification methods in various ecology studies

The more applied an ecological study is, the more the emphasis is on the effects on ecological communities of particular environmental factors, for example pollutants, management regimes, and other human-induced changes in the environments. (ter Braak, 1994).

Correspondingly, the statistical analysis should not 'just' show the major variation in the species assemblage, but focus on the effects on the variables of prime interest. Applied studies thus call for direct methods of ordination, typically with a very limited number of (qualitative or quantitative) environmental variables (ter Braak, 1994)

The range of community variation in an applied study tends to be quite small compared to that in the early ordination studies (e.g. Whittaker, 1965; Hill & Gauch, 1980).

Determining which factors control the distribution patterns of plant communities remains a central goal in ecology Classification assumes from the outset that the species assemblages fall into discontinuous groups, whereas ordination starts from the idea that such assemblages vary gradually.

Ordination compares sites on their degree of similarity, and then plots them in Euclidian space, with the distance between points representing their degree of Similarity. Ordination techniques include principal components analysis (PCA), detrended correspondence analyses (DCA), and nonmetric multidimensional scaling (NMS)

Ordination (or inertia) methods, like principal component and correspondence analysis, and clustering and classification methods are currently used in many ecological studies (Zare Chahouki et al., 2009, Anderson, 2002; Gauch et al., 1977; Orloci, 1975; Whittaker (ed.), 1967; Legendre & Legendre, 1988).

Ordination methods can be divided in two main groups, direct and indirect methods. Direct methods use species and environment data in a single, integrated analysis. Indirect methods use the species data only (Jongman et al., 1987). In contrast, if a unimodal response model is assumed, the relationships are unimodal. Unimodal relations are usefully summarized by their modes or - more conveniently - weighted averages (ter Braak & Looman, 1986), so that a sensible coefficient for the species × environment table is the weighted average. The other way round, if a correlation coefficient is chosen, the implied response model is linear or approximately linear and if the chosen coefficient is the weighted average than the implied response model is unimodal (i.e. if the true model is bimodal, the ordination will fail, and if the true model is linear, the ordination will be inefficient). Assumptions about the response model and the choice of coefficients to use in secondary tables are thus interrelated.

New methods of exploring differences among groups include the nonparametric, recursive classification, and regression tree (CART). It is used to classify habitats or vegetation types and their environmental variables (McCune et al., 2002). It produces a top-to-bottom visual classification tree that undergoes a "pruning" or optimization process. CART is used to generate community maps, wildlife habitats, and land cover types in conjunction with a GIS (see p. 195 in Chapter 3). Another multivariate technique is Structural Equation Modeling (unfortunately termed SEM), a merger of factor analysis and path analysis (McCune et al. (2002). It is a method of evaluating complex hypotheses (e.g., effects of abiotic factors on plant species richness).

With multiple causal pathways among variables. It requires the initial development of a path diagram. It is an analysis of covariance relationships, effectively limited to about 10 variables. See Shipley (2000), McCune et al. (2002), Pugesek et al. (2002) and Bakus (2007).

7. References

Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. Austral. Ecol. 26:32-46.

Alisauskas, R.T. (1998). Winter range expansion and relationships between landscape and morphometrics of midcontinent Lesser Snow Geese. Auk 115(4):851-862.

Anderson, C.W., Barnett, V., Chatwin, P.C. and El-Shaarawi A.H. (2002). Quantitative Methods for Current EnvironmentalIssues. Springer-Verlag, New York.

C.T. Bastian, S.R. Koontz and Menkhaus D.J. (2001). The Impact of Forward Contract Information on the Fed Cattle Market: An Experimental Investigation into Mandatory Price Reporting, UW Agricultural and Applied Economics Seminar. August 31, 2001. (Presented by Bastian).

Bakus Gerald J. (2007). Quantitative Analysis of Marine Biological Communities Field Biology and Environment. WILEY-INTERSCIENCE, A John Wiley & Sons, Inc., Publication, 453p.

Beals, E.W. (1973). Ordination: Mathematical elegance and ecological naivete. J. Ecol. 61:23–35.

Bray, J.R. and Curtis J.T. (1957). An ordination of the upland forest communities of southern Wisconsin.Ecol. Monogr. 27:325-349.

Carignan, V. and Villard M. (2002). Selecting indicator species to monitor ecological integrity: a review. Environ. Monitor. Assess. 78: 45-61.

Chytr, M. (2002). Determination of diagnostic species with statistical fidelity measures. J. Veg. Sci. 13: 79–90.

Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol.18:117-143.

Clarke, K.R. and Warwick R.M. (2001). Change in Marine Communities: An Approach to Statistical Analysis and Interpretation. 2nd edition. PRIMER- E, Plymouth Marine Laboratory, Plymouth, U.K.

Clifford, H.T. and Stephenson W. (1975). An introduction to numerical classification. Academic Press, New York, pp. 229.

Cochran, W.G. (1977). Sampling Techniques. Wiley, New York.

Davis, J.C. (1986). Statistics and Data Analysis in Geology. Wiley, New York.

Diserud, O.H. and Aagaard K. (2002). Testing for changes in community structure based on repeated sampling. Ecology, 83(8): 2271–2277.

Dufrene, M. and Legendre P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67:345–366.

Dunn, C. P., and Stearns F. (1987). Relationship of vegetation layers to soils in southeastern Wisconsin forested wetlands. Am. Midl. Nat. 118:366-74.

Dytham, C. (1999). Choosing and Using Statistics: A Biologist's Guide. Blackwell Publishing, Williston, VT.

Ferson, S., Downey P., Klerks P., Weissburg M., Kroot S.I., S. Jacquez O., Ssemakula J., Malenky R. and Anderson K. (1986). Competing reviews, or why do Connell and Schoener disagree? Am. Nat., 127: 571–576.

Feoli, E., and Orl¢ci L. (1992). Thre properties and interpretation of observations in vegetation study. Coenoses 6:61-70.

Fortin, M-J. and Gurevitch J. (2001). Mantel tests: Spatial structure in field experiments. pp. 308-326 in: Scheiner S.M. and J. Gurevitch (eds.). Design and Analysis of Ecological Experiments. Oxford University Press, Oxford.

Gauch, H.G. (1977). ORDIFLIX—A flexible computer program for four ordination techniques: weighted averages, polar ordination, principal components analysis, and reciprocal averaging. In: Ecology and Systematics, Cornell University, Ithaca, N.Y.

Gauch, H.G., J. (1982). Multivariate Analysis in Community Ecology. Cambridge University Press, New York.

Gittins, R. (1985). Canonical analysis: a review with applications in ecology. Springer-Verlag, Berlin.

Gotelli, N.J. and Ellison A.M. (2004). A Primer of Ecological Statistics. Sinauer Associates, Sunderland, Maine

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis. Biometrika 53, 325-338.

Green, R.H. (1979). Sampling Design and Statistical Methods for Environmental Biologists. John Wiley and Sons, New York.

Greig-Smith, P. (1983). Quantitative Plant Ecology, 3rd Edition. Blackwell Scientific Publications, London, 359 pp.

Green, R.H. 1989. Power analysis and practical strategies for environmental monitoring. Environ. Res. 50:195–205.

Hill, M.O., Bunce R.G.H. & Shaw M.V. (1975). Indicator species analysis, a divisive polythetic method of classification, and its application to survey of native pinewoods in Scotland. Journal of Ecology, 63: 597–613

Hill, M.O. (1979). TWINSPAN – a FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes. Ithaca: Section of Ecology and Systematic, Cornell University.

Hill M.O. & Gauch H.G. (1980). Detrended correspondence analysiss, an improved ordination technique. Vegetatio, 42: 47-58

Jaccard, P. (1901): Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Socie´ te´ Vaudoisedes SciencesNaturelles, 37: 547–579.

Jongman, R.H.G., ter Braak C.J.F. and Van Tongeren O.F.R. (1987). Data analysis in community and landscape ecology. Cambridge University Press, Cambridge, UK.

Jongman, R.H.G., Ter Braak, C.J.F. and van Tongeren O.F.R. (eds.) (1995). Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

Kent, M. (2006). Numerical classification and ordination methods in biogeography. Progress in Physical Geography, 30: 399-408

Krebs, C.J. (1989). Ecol. Method. Harper Collins, New York.

Krebs, C.J. (1999). Ecological Methodology. Harper & Row, New York.

Kutner, M.H., Nachtscheim, C.J., Wasserman, W. and Neter J. (1996). Applied linear statistical models. WCB/McGraw-Hill, New York.

Lee, Y.W. and Sampson D.B. (2000). Spatial and temporal stability of commercial ground fish assemblages off Oregon andWashington as inferred from Oregon travel logbooks. Canadian J. Fish. Aqua. Sci., 57:2443-2454.

Legendre, P. & Gallagher E.D. (2001). Ecologically meaningful transformations for ordination of species data. Oecologia, 129: 271–280

Legendre, P. and Legendre L. (1998). Numerical Ecology. 2nd Edition.Elsevier, Amsterdam.

Lepš, Jan & Šmilauer P. (1999). Multivariate Analysis of Ecological Data Faculty of Biological Sciences, University of South Bohemia Ceské Budejovice,110pp

Ludwig, J.A., Reynold, J.F. (1988). Statistical Ecology. Wiley, New York, 337pp

Niemi, G.J. and McDonald M.E. (2004). Application of ecological indicators. – Annu. Rev. Ecol. Evol. Syst. 35: 89–111.

Manly, B.F.G. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman and Hall, London.

Mantel, N. 1967. The detection of disease clustering and generalized regression approach. Cancer Res. 27:209-220.

Mantel, N. and Valand R.S. (1970). A technique of nonparametric multivariate analysis. Biometrics 26:547–558.

McCune, B. and Mefford M.J. (1999). PCORD, Multivariate Analysis of Ecological Data, Version 4. MjM Software Design, Gleneden Beach, Oregon, USA.

McCune, B., Grace J.B. and Urban D.L. (2002). Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon.

McGarigal, K., Cushman, S. and Stafford S. (2000). Multivariate Statistics for Wildlife and Ecology Research. Springer-Verlag, New York

McGeogh, M.A. (1998). The selection, testing and application of terrestrial insects as bioindicators. Biol. Rev. 73: 181–201.

McGeoch, M.A. and Chown. S.L. (1998). Scaling up the value of bioindicators. Trends Ecol. Evol. 13: 46-47.

Mielke, P.W. & Berry K.J. (2001). Permutation methods: A distance function approach. New York: Springer-Verlag.

Mielke, P. W., Berry K.J. & Johnson E. S. (1976). Multi-response permutation procedures for a priori classications. Communications in Statistics- Theory and Methods, 5: 1409-1424

Mitchell, R.J. (2001). Path analysis: Pollination. pp. 217–234 in: Scheiner S.M. and J. Gurevitch (eds.). Design and Analysis of Ecological Experiments. Oxford University Press.

Petratis, P.S., Beaupre S.J. and Dunham A.E. (2001). ANCOVA: Nonparametric and Randomization Approaches. pp. 116–133 in: Scheiner S.M. and Gurevitch (eds.). Design and Analysis of Ecological Experiments. Oxford University Press, Oxford.

Pugesek, B., Tomer A. and von Eye A. (eds.) (2002). Structural Equations Modeling: Applications in Ecological and Evolutionary Biology Research. Cambridge University Press, Cambridge, U.K.

Paukert, C.P. and Wittig T.A. (2002). Applications of multivariate statistical methods in fi sheries. Fisheries 27(9):16-22

Palmer, M.W. (1988). Fractal geometry: a tool for describing spatial patterns of plant communities. Vegetatio 75:91–102.

Odum, E.P. and Barrett G.W. (2005). Fundamentals of Ecology. Thomas Brooks/Cole, Belmont, CA.

Orloci, L. (1975). Multivariate Analysis inVegetation Research. Junk, The Hague.

Rohlf, F.J. (1995). BIOM: A Package of Statistical Programs to Accompany the Text Biometry. Exeter Software, Setauket, New York.

Scheiner, S.M. (2001). Theories, hypotheses, and statistics. pp. 3-13 in: Scheiner S.M. and J. Gurevitch (eds.). Design and Analysis of Ecological Experiments. Oxford University Press, Oxford.

Shipley, B. (2000). Cause and Correlation in Biology. Cambridge University Press, Cambridge, U.K.

Statsoft, Inc. 1995. STATISTICA for Windows. 2nd edition. Tulsa, OK.

Stiling, P. (2002). Ecology: Theories and Applications. Prentice Hall, Upper Saddle River, N.J.

Swan, J.M.A. (1970). An examination of some ordination problems by use of simulatedvegetational data. Ecology 51: 89–102.

Ter Braak C.J.F. & Looman C.W.N. (1986). Weighted averaging, logistic regression and the Gaussian response model. Vegetatio, 65: 3-11

Ter Braak, C.J.F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio, 69:69-77.

Ter Braak, C. J. F. (1994). Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience 1 (2), 127-140.

Ter Braak C.J.F. & Šmilauer P. (1998). CANOCO Reference Manual and User's Guide to Canoco for Windows. Microcomputer Power, Ithaca, USA. 352 pp.

Thompson, S.K. (2002). Sampling. John Wiley & Sons, New York. Second Edition.

Turner MG. (1989). Landscape ecology: the effect of pattern on process. Ann. Rev. Ecol. Syst. 20:171-197.

Turner MG, RH Gardner and RV O'Neill (2001). Landscape Ecology in Theory and Practice: Pattern and Process. Springer, New York.

van der Meer, J. Heip C.H., Herman P.J.M., Moens T. and van Oevelen D. (2005). Measuring the Flow of Energy and Matter in Marine Benthic Animal Populations. pp. 326–408 in: Eleftheriou. A. and A. McIntyre (eds.). 2005. Methods for Study of Marine Benthos. Blackwell Science Ltd., Oxford, UK.

Van der Plogeg, S.W.F. & Vlijm L. (1978). Ecological evaluation, nature conservation and land use planning with particular reference to the methods used in the Netherlands.Biol.Consero.14:197-221.

Van Sickle, J. (1997). Using mean similarity dendrograms to evaluate classifications. Journal of Agricultural, Biological, and Environmental Statistics, 2:370-388.

Whittaker R.H. (1965). Dominance and diversity in land plant communities. Science 147: 250–260.

Whittaker, R.H. (1967). Gradient analysis of vegetation. Biol. Rev. 42:207-264.

Zare Chahouki, M. A. Azarnivand H., Jafari M. & Tavili A. (2009). Multivariate Statistical Methods as a Tool for Model_Based Prediction of Vegetation Types, Russian Journal of Ecology, 41(1): 84-94.

Zare Chahouki, M.A. (2006). Modeling the spatial distribution of plant species in arid and semi-arid rangelands. PhD Thesis in Range management, Faculty of Natural Resources, University of Tehran, 180 p. (In Persian).

Schluter. D. and Grant P.R. (1982). The distribution of *Geospiza difficilis* on Galapagos islands: test of three hypotheses. Evolution 36:1213-1226

Earth and Environmental Sciences
Edited by Dr. Imran Ahmad Dar

ISBN 978-953-307-468-9
Hard cover, 630 pages
Publisher InTech
Published online 07, December, 2011
Published in print edition December, 2011

We are increasingly faced with environmental problems and required to make important decisions. In many cases an understanding of one or more geologic processes is essential to finding the appropriate solution. Earth and Environmental Sciences are by their very nature a dynamic field in which new issues continue to arise and old ones often evolve. The principal aim of this book is to present the reader with a broad overview of Earth and Environmental Sciences. Hopefully, this recent research will provide the reader with a useful foundation for discussing and evaluating specific environmental issues, as well as for developing ideas for problem solving. The book has been divided into nine sections; Geology, Geochemistry, Seismology, Hydrology, Hydrogeology, Mineralogy, Soil, Remote Sensing and Environmental Sciences.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821