Weed Science, 50:281–292. 2002

Review

Multivariate analysis in weed science research

N. C. Kenkel
Corresponding author. Department of Botany, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2; kenkel@cc.umanitoba.ca

D. A. Derksen
Agriculture and Agri-Food Canada, P.O. Box 1000F, R.R. #3, Brandon, Manitoba, Canada R7A 5Y3

A. G. Thomas
Agriculture Canada Research Station, 107 Science Cres., Saskatoon, Saskatchewan, Canada S7N 0X2

P. R. Watson
Agriculture and Agri-Food Canada, P.O. Box 1000F, R.R. #3, Brandon, Manitoba, Canada R7A 5Y3

Data containing many variables are often collected in weed science research, but until recently few weed scientists have used multivariate statistical methods to examine such data. Multivariate analysis can be used for both descriptive and predictive modeling. This paper provides an intuitive geometric introduction to the more commonly used and relevant multivariate methods in weed science research, including ordination, discriminant analysis, and canonical analysis. These methods are illustrated using a simple artificial data set consisting of abundance measures of six weed species and two soil variables over 12 sample plots.

Key words: Canonical correlation analysis (CANCOR), canonical correspondence analysis (CCA), canonical discriminant analysis (CDA), correspondence analysis (CA), nonmetric multidimensional scaling (NMDS), ordination analysis, principal component analysis (PCA), principal coordinate analysis (PCoA), redundancy analysis (RDA), statistical methods.

Multivariate ("many variable") statistical techniques are powerful tools for investigating and summarizing underlying trends in complex data structures (Legendre and Legendre 1998). The term "multivariate" refers to the methods that undertake a simultaneous analysis of several variables or dimensions. Multivariate statistical methods were developed in the early 20th century, but computational limitations delayed their wider application until the advent of high-speed computers. In biology, multivariate methods were first used by plant ecologists to explore and model vegetation survey data (Goodall 1954; Orlóci 1966). More recently, agricultural scientists have used these scaling methods to analyze and interpret complex survey and experimental data (e.g., Derksen et al. 1995; Post 1988; Thomas and Frick 1993). Although multivariable data are often collected in weed science and related disciplines, the potential advantages of a multivariate approach to analyzing such data are not always appreciated. Proper application and interpretation of multivariate techniques requires an understanding of the theory, assumptions, and limitations of the various methods available.

In this paper we present a selective, nonstatistical overview of the more commonly used multivariate scaling methods, including descriptive ordination models and predictive canonical models. Our objective is to instill in the reader an intuitive understanding of these methods and their applications using simple numerical examples. Although some basic computational aspects are discussed, the reader is referred to specialized monographs for more detailed information on the underlying theory and mathematics.

What are Multivariate Data?

Multivariate data arise when attributes for more than one variable are measured on each sampling unit within the context of a sample survey or experiment. The resulting data are summarized in matrix form as a set of p variables measured on each of the n sampling units (Figure 1). An example is a data set consisting of density values for p weed species (variables) enumerated in each of the n field plots (sampling units). An interpretative analysis of multivariate data considers all the variables simultaneously rather than as a set of p independent variables. Coordinated responses among variables result in an underlying structure to multivariate data. A primary objective of multivariate analysis is to detect and effectively summarize this underlying structure. A univariate approach, in which each variable is examined independently of the others, will fail to detect the higher-order responses that define the underlying community structure.

Objectives of Multivariate Analysis

Following Jeffers (1982), we define a model as a formal expression of the relationship between sampling units expressed in mathematical terms. A number of modeling approaches are used in agricultural sciences, including dynamic, stochastic, optimization, game theory, catastrophe theory, matrix, and multivariate modeling. Two major objectives of multivariate modeling can be distinguished: descriptive modeling that typically involves summarizing underlying data structure and predictive or confirmatory modeling that generally involves statistical hypothesis testing (Jeffers 1988). Descriptive modeling, a common objective when analyzing sample survey data, is undertaken to achieve a parsimonious representation of the underlying data structure and to summarize variable intercorrelations. This approach is lucidly described by Rao (1964):

> When a large number of measurements [variables] are available, it is natural to enquire whether they could be replaced by a fewer number of the measurements or of

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3n} \\ . & . & . & \cdots & . \\ X_{p1} & X_{p2} & X_{p3} & \cdots & X_{pn} \end{bmatrix}$$

FIGURE 1. Generalized matrix **X** of p variables (rows) and n sampling units (columns). An element of the matrix (X_{ij}) denotes the value of the ith variable in the jth sampling unit.

their functions, *without loss of much information,* for convenience in the analysis and in the interpretation of data.

In analyzing sample survey data, multivariate methods are generally used as exploratory descriptive modeling tools for generating hypotheses regarding the causal mechanisms producing an underlying data structure. By contrast, statistical hypothesis testing is a confirmatory or predictive approach undertaken within the context of more formalized survey and experimental designs. In both descriptive and predictive modeling, data reduction and summarization are the distinguishing features of multivariate analysis (Legendre and Legendre 1998).

Data Structures and Data Partitioning

Two basic multivariate data structures are distinguished: biotic data comprised response variables, and abiotic data comprised factor variables (Figure 2). As an example of a biotic data set, consider again a survey of weed species abundance across a series of field plots. All the variables quantified are measured on the same scale, i.e., weed density by species. A biotic data set typically has two distinguishing features: a high proportion of zeros because many species are absent from most fields and occasional "hot spots" where a given species is locally abundant (e.g., species *F* in Plot 3, Figure 2). These features render many biotic data sets nonlinear, which in turn compromises the effectiveness of linear multivariate methods (Kenkel and Orlóci 1986; Legendre and Legendre 1998). As an example of an abiotic data structure, consider measurements of soil variables in the same field plots. In this case the variables quantified are often measured on different scales (e.g., pH, nutrient concentration, percent organic matter), making it necessary to standardize the variables to render them commensurable. Because most abiotic data have no zero entries and lack hot spots, they produce broadly linear data structures that are most effectively summarized using linear multivariate methods.

Multivariate data matrices can be partitioned in various ways (Figure 3). Unpartitioned data are analyzed using standard ordination methods, with the objective of exploring and summarizing underlying trends. More sophisticated multivariate methods are required when the data have an underlying structure related to a specific sampling or experimental design. Canonical discriminant analysis (CDA) is appropriate when the objective is to examine relationships among sampling units naturally partitioned into two or more groups or treatments, whereas canonical analysis is used to examine relationships between two variable sets (factor and response variables) measured on the same sampling units. More complex partitionings are of course possible; for example, repeated measures add a third dimension to the basic data set (species by plot by time). Complex designs present considerable challenges to the analysis and interpretation of multivariate data structures (Green 1979, 1993).

Variable Selection, Standardization, and Transformation

The proper selection of variables is critical to the success of any multivariate survey or experiment. A clear statement of study objectives helps ensure that the variables selected are relevant to the task at hand. Jeffers (1988) emphasizes this point:

> Much time and effort have been wasted by the application of multivariate analysis to sets of data containing a "rag-bag" of variates included because they were easy to measure, or because they happened to be available, without any apparent consideration of the logical design of the investigation.

Abiotic variables should be selected with care to ensure that they are pertinent to the specific objectives of the study. Selection of biotic variables is less problematic because researchers generally collect information on all species encountered during a study. However, rare species may be eliminated from the data set before multivariate analysis because they may unduly influence the results.

The most commonly used multivariate methods assume ratio variables (quantitative variables with a defined zero),

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
A	1	3	3	4	7	9	3	5	7	9	10	12
B	4	5	3	7	8	11	1	2	3	6	8	9
C	5	8	0	0	0	0	0	3	0	0	0	0
D	2	7	0	0	3	1	7	0	0	0	0	0
E	1	0	0	0	7	10	0	1	0	0	7	9
F	0	0	8	6	0	0	0	0	4	0	0	0
N	5	5	8	3	7	8	3	3	2	5	6	10
K	7	5	9	2	6	6	5	3	1	3	4	7

FIGURE 2. The data set used in the examples. Variables *A*–*F* are the abundance of six weed species (response variables, biotic data set), whereas *N* and *K* are two soil variables (factor variables, abiotic data set). The 12 sampling units (columns 1 to 12) fall into two groups (1 to 6 and 7 to 12).

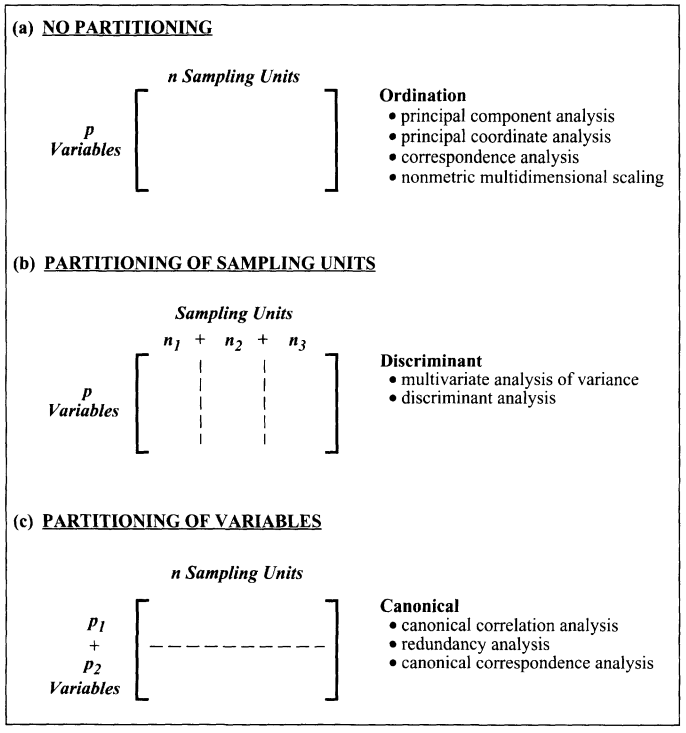

FIGURE 3. Data sets, partitionings, and methods appropriate to their analysis: (a) no partitioning, (b) partitioning of sampling units, (c) partitioning of variables. Modified from Green 1979.

of which counts, densities, and lengths are examples. Specialized methods and approaches are required for mixed data sets that contain nominal (unordered) or ordinal (rank order) variables (Orlóci 1978). Interval variables (quantitative variables lacking a defined zero, e.g., compass direction) should be recoded as ratio variables before analysis.

Standardization is required to render variables commensurable when they are measured on different scales. The most common is the "unit-variance" standardization, which involves expressing a measured value as the deviation from the variable mean followed by division by the variable's standard deviation. This standardization places all the variables on a common scale; each is standardized to zero mean and unit variance, rendering them additive and dimensionless (Jeffers 1988). The unit-variance standardization is used in most multivariate analyses and is implicitly incorporated in the familiar product–moment correlation coefficient. Sampling units may also be standardized. For example, species abundance data may be standardized to a sum to 100% in each sample plot. This transformation is appropriate if proportional rather than absolute differences in species abundance among plots are of interest.

Variable transformations are undertaken to ensure that variables conform to an assumed underlying distribution (e.g., multivariate normality) to increase the linear relationship among variables and to reduce heteroscedasticity of the variance–covariance structure. Many multivariate methods assume an underlying normal (or multivariate normal) distribution, although strict adherence to this assumption is required only when undertaking formal hypothesis testing. Normalizing transformations have the added benefit of reducing the influence of data outliers to which most multivariate methods are sensitive. The logarithmic transformation renders a multiplicative series additive, making it appropriate in cases where a biological population (e.g., a weed newly introduced into a field) is capable of exponential increases in cover abundance. A square root transformation is often appropriate for count data that follow the Poisson distribution, whereas the arcsine transformation is generally applied to proportional data. Choice of an appropriate transformation is dependent on the structure of the data being analyzed, the objectives of the investigation, and the multivariate methods used (Zar 1974).

Selection and Enumeration of Sampling Units

Taken together, the sampling units enumerated in a multivariate survey or experiment constitute a sample from a much larger population. Statistical inference requires that the sample be representative, which can be accomplished using randomization methods developed by sampling theorists (Cochran 1977). The analysis of multivariate data from agroecosystem experiments raises additional issues that are critical to the proper interpretation of results, including data collection, nature of system-level treatments, and experimental design. Consistency in data collection is critically important in ensuring that results are comparable in space (between sites) and over time (between years). Ideally, the same number of sampling units should be enumerated at each site and at each date, and the data should be collected at the same time of the year or at the same stage of crop development (Derksen et al. 1998). Because weed communities are characterized by fluctuational rather than directional dynamics (Derksen et al. 1995), weed communities should be enumerated over many years. If data are collected only at the end of a multiyear trial, cumulative treatment effects and growing conditions at the time of sampling will be confounded and difficult to separate.

Although treatment combinations in weed agroecosystem experiments often have simple labels (e.g., crop rotation, low-input treatment, tillage system), they in fact represent composites of numerous interacting factors. For example, the effect of different crop rotations on weed community composition is the result of crop competitive ability coupled with crop management practices, such as seeding date, seeding rate, row spacing, fertilizer timing and placement, herbicide usage, and interrow cultivation. Results should, therefore, be interpreted as system rather than simple treatment effects (Jeffers 1978). Indeed, a clear statement of system effects is critical if results from different research programs are to be compared.

A balanced experimental design is important when comparing weed management strategies and weed communities across agronomic systems. Traditional herbicide experiments have weed-free or weed-present controls, whereas in systems-level experiments one compares the relative effects of differing systems on the weed community. For example, the appropriate "check" for a low-input system would be a conventional-input system, each employing the same herbicides, seeding dates, and so forth. Balance is achieved by having sufficient commonality such that system effects can be parsed out, even if treatment details are confounded. For example, an experiment comparing input levels may use different herbicides and seeding dates but may include the same crops grown in rotation. Weed management strategies should be designed by taking into consideration the trial objectives and the system definitions. Trials that utilize non-

registered herbicides, multiple applications of herbicides, tillage to ensure weed-free conditions, or rotations atypical of the systems being compared (e.g., organic farming vs. zero tillage) may meet agronomic objectives but may not be appropriate for weed community comparisons. It is also critical that all phases of crop rotations be present in each year to avoid confounding system effects and year-to-year climatic variability (Derksen et al. 1995).

Descriptive Multivariate Modeling: Ordination Methods

Multivariate data are structurally complex, making preliminary exploratory analysis of underlying data structures critical to the success of any descriptive or predictive modeling program. Exploratory analysis should be viewed as a highly flexible exercise designed to elucidate and summarize underlying trends in the data (Tukey 1977). Such analyses are preliminary in the sense that they provide insights into the data structure, which in turn direct the user to more meaningful analyses and interpretations. An optimal exploratory analysis is achieved using a graphical interface that dynamically produces scatterplots, frequency histograms, and various summary statistics of the variables and sampling units (Chambers et al. 1983). Exploratory analysis is well suited to the detection of outliers and nonlinear trends, allowing the user to make informed decisions in choosing the most appropriate data transformation and multivariate approach.

It is recommended that multivariate modeling proceed in steps, carefully examining the results obtained at a given stage before proceeding further (Jeffers 1988). Indeed, experienced multivariate data analysts view modeling as an adaptive learning process, in which decisions made at each stage of the analysis direct subsequent steps and strategies (Tukey 1977).

Ordination refers to a group of analytical methods designed to represent a complex multivariate data structure in a low-dimensional space, while retaining as much of the underlying trended variation as possible. This objective is achievable provided that the variables are intercorrelated or trended with one another. Ordination methods are generally used in an exploratory strategy to search for and summarize underlying trends. This approach is often referred to as indirect gradient analysis (ter Braak and Prentice 1988) because ordination is used to generate hypotheses regarding the causal factors underlying the trends present in the data. A number of ordination or scaling methods are available, including principal component analysis (PCA), principal coordinate analysis (PCoA), correspondence analysis (CA), and nonmetric multidimensional scaling (NMDS). Selection of the appropriate ordination method is dependent on the underlying data structure as well as on the study objectives.

Principal Component Analysis

In PCA a p-dimensional data space is represented as a set of mutually perpendicular (orthogonal) ordination axes. Successive ordination axes are obtained through rigid rotation of the original p variable axes to maximize the linear variation accounted for. As a simple example, consider a scattergram of 12 field plots with respect to $p = 2$ variables (species A and B, Figure 2). The example is very simple (normally $p > 2$) but useful for illustrative purposes. The 12 plots are represented as points in two-dimensional species space (Figure 4) together with three sets of axes: (1) the original variable axes, labeled species A and B; (2) the same

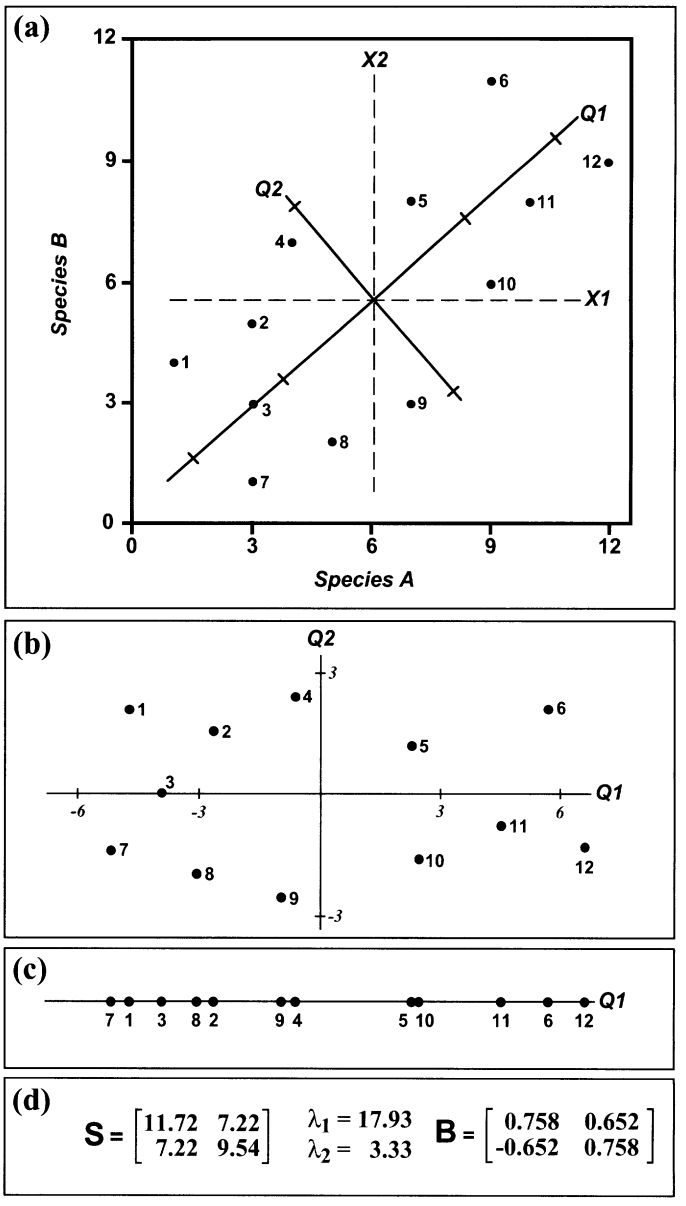

FIGURE 4. Principal component analysis (PCA) of 12 sampling units (1 to 12), using data from species A and B in Figure 2. (a) Scattergram of the 12 sampling units in two-dimensional species space, showing the original species axes centered on their respective means (dashed lines $X1$ and $X2$) and the two-fitted component axes $Q1$ and $Q2$. (b) Two-dimensional ordination of the 12 sampling units, which in this example retains 100% of the total variance. (c) A reduced (one dimensional) ordination of the data on the first (principal) component axis $Q1$. This representation retains $\lambda_1 / \Sigma \lambda_t = 17.93/21.26 = 84.34\%$ of the total variance. (d) Covariance matrix **S**, eigenvalues λ, and eigenvector matrix **B** of the example data set. In the covariance matrix, the diagonal elements $S_{11} = 11.72$ and $S_{22} = 9.54$ are the variances of species A and B, respectively; the off-diagonal element $S_{12} = S_{21} = 7.22$ is the covariance of species A and B (covariance is positive because the two species are positively correlated). The principal eigenvalue $\lambda_1 = 17.93$ is the variance of sampling unit scores along the first axis (refer to Panel c). Note that $\lambda_1 + \lambda_2 = S_{11} + S_{22}$, indicating that PCA is a simple repartitioning of the total variance. Matrix **B** contains eigenvector elements: for example, $b_{11} = 0.758$ is the direction cosine of species A and component axis $Q1$, $b_{12} = 0.652$ is the direction cosine of species B and component axis $Q1$.

variable axes, but centered on their respective means (i.e., the centroid of the plot swarm), labeled *X1* and *X2;* and (3) the ordination or principal component axes, labeled *Q1* and *Q2.* Note that the relative plot positions are not changed; *Q1* and *Q2* simply describe a new coordinate system for the 12 plots, each based on a linear combination of species *A* and *B.* The first ordination axis (*Q1*) represents a specific rotation of axes *X1* and *X2* such that linear variation along *Q1* is maximized. This axis is uniquely defined; it minimizes the sum of squares perpendicular point-to-line distances, which is equivalent to maximizing the variance along *Q1.* The variance of the plot coordinates (ordination scores) on *Q1* is termed the eigenvalue λ_1. The second ordination axis (*Q2*) must be orthogonal (perpendicular) to *Q1.* In our simple two-dimensional example *Q2* is automatically defined once *Q1* has been located, but this is not the case for higher-dimensional ($p > 2$) data. Specifically, when $p > 2$ the second axis *Q2* must be orthogonal to *Q1* while at the same time accounting for the maximum amount of residual linear variation not already accounted for by *Q1.* For multivariate data, third and subsequent axes are defined in the same way. The PCA solution for a given data set is, therefore, entirely deterministic.

It is important to recognize that PCA does not reduce dimensionality per se (Figure 4). Instead, dimension reduction is achieved by ignoring higher-ordination (numerically unimportant) axes and displaying the point configuration on the first few, most highly trended axes (Figure 4c). If the data have a strong underlying linear structure, the first few ordination axes will provide a reasonable and parsimonious representation of major trends underlying the data. For the two-dimensional case, the analogy to simple regression analysis is apparent. In regression analysis the relationship between two variables is expressed unidimensionally, i.e., as a "line of best fit" that is similar (though not identical) to ordination axis *Q1.* Methods for determining the number of statistically significant ordination axes are summarized in Jackson (1993) and Legendre and Legendre (1998).

In practice principal component axes are obtained through eigenanalysis of a *p*-dimensional covariance (or correlation) matrix **S.** Because a covariance matrix summarizes only linear relationships, PCA is not suitable for the analysis of nonlinear data structures. For a given variance–covariance matrix **S,** unique sets of eigenvalues and eigenvectors are sought that satisfy the matrix equation:

$$[\mathbf{S} - \boldsymbol{\lambda}\mathbf{I}]\mathbf{B} = 0 \qquad [1]$$

where $\boldsymbol{\lambda}$ is a row vector of *p* eigenvalues, **B** is a matrix of *p* eigenvectors, and **I** is the identity matrix (a square matrix containing ones as the diagonal elements, and zeroes as the off-diagonal elements). The sum of eigenvalues equals the sum of the individual variances of the *p* variables because PCA is simply a linear transformation that repartitions the total variance along linearly orthogonal component axes. The extracted component axes are, therefore, linear combinations of the original variables. Specifically, each variable is "weighted" on each component axis, according to its correlation with the trend summarized by that axis. Variable weights on the component axes are contained in the eigenvector matrix **B.** An element b_{ij} of **B** is a direction cosine (range −1 to +1) or the cosine of the angle between variable axis *i* and the *j*th component axis. The larger the absolute value of the direction cosine, the more highly weighted the variable is on the ordination axis. The product–moment correlation of the *i*th variable and the *j*th component axis is:

$$r_{ij} = b_{ij}\sqrt{[\lambda_i/S_i^2]} \qquad [2]$$

where S_i^2 is the variance of the ith variable. Direction cosines are critical to the interpretation of PCA results because they indicate the relative contribution of each variable to the major structural trends (ordination axes) underlying the data. Direction cosines can be summarized in the tabular form, or alternatively they can be graphed directly on the ordination diagram to produce a biplot (Gabriel 1981; Podani 1994).

The coordinate positions or component scores of the *n* sampling units on the derived component axes are simple linear combinations of the variable weights (direction cosines). For example, the score of individual X_j on component axis 1 is:

$$Q_{1j} = b_{11}(X_{1j} - \bar{X}_1) + b_{12}(X_{2j} - \bar{X}_2) + \cdots + b_{1p}(X_{pj} - \bar{X}_p) \qquad [3]$$

A given component score Q_{ij} is thus defined as a linear combination of the *p* variables, with the direction cosines as multiplying factors.

In the earlier presentation, it is assumed that **S** is a covariance matrix. The use of covariance in PCA is appropriate when variables are measured on the same scale (e.g., biotic data sets). If the variables are measured on different scales (e.g., many abiotic data sets), a correlation matrix must be specified to render the variables commensurable. Because each of the *p* variables is scaled to unit variance, the eigenvalues from a PCA of a correlation matrix sum to *p.*

In weed science research, PCA is the method of choice when the underlying data structure is broadly linear (i.e., relatively few zero values in the data set). Example applications include Légère and Samson (1999) and Ominski et al. (1999).

Principal Coordinate Analysis

PCoA, which is also known as metric multidimensional scaling, produces a map of the *n* individuals in ordination space such that the pairwise distances among individuals in that space match as closely as possible the corresponding distances in variable space (compare with NMDS, which is discussed subsequently). PCoA can also be viewed as a generalized variant of PCA because the method exploits the close relationship between covariance and Euclidean metric spaces (Gower 1966) to perform an eigenanalysis of an *n*-dimensional distance matrix (Digby and Kempton 1987). In fact, for a given data set, PCoA of the *n*-dimensional Euclidean distance matrix produces an ordination identical to that obtained from PCA of a *p*-dimensional covariance matrix. The chief advantage of PCoA lies in its generality because the method can be used to perform eigenanalysis on a wide variety of broadly metric distance measures (Legendre and Legendre 1998). A drawback of the method is that variable direction cosines and biplot scores are not available because it operates directly in *n*-dimensional space. Although PCoA has not been widely used, it has considerable

potential in analyses where covariance–correlation calculations are inappropriate or ill defined (Jeffers 1988).

Correspondence Analysis (CA)

CA refers to a family of closely related ordination methods that includes reciprocal-weighted (two way) averaging (Hill 1973; ter Braak 1995), dual scaling, and canonical analysis of contingency tables (Greenacre 1984). Despite considerable differences in statistical derivation, these various CA algorithms produce identical results, except for minor variations in the scaling of ordination scores (Benzecri 1992; Greenacre 1984). CA can also be viewed as a special case of PCoA, in which the data are doubly standardized by the row and column totals (Digby and Kempton 1987, page 90).

One approach is to consider CA as a multivariate contingency table eigenanalysis of a data matrix **X.** The total inertia of **X** is $\chi^2/X_{..}$ where χ^2 is the familiar chi-squared statistic:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{n} [X_{ij} - E_{ij}]/E_{ij} \quad [4]$$

and $E_{ij} = [X_{i.}X_{.j}/X_{..}]$, $X_{i.}$ is the ith row total, $X_{.j}$ is the jth column total, and $X_{..}$ is the grand total.

Inertia measures the degree of correspondence between the row and column categories (variables and sampling units, respectively) of the matrix **X.** In CA eigenanalysis is used to partition the total inertia into k = MIN (p,n) linear additive components (Digby and Kempton 1987):

$$\chi^2 = \sum_{i=1}^{k} \chi^2_i = X_{..} \sum_{i=1}^{k} R_i^2 \quad [5]$$

The eigenvalue $\lambda_i = R_i^2$ associated with the ith ordination axis is a squared canonical correlation (range 0 to 1), which is interpretable as a measure of the correspondence between variables and sampling units. CA is, therefore, a special case of canonical correlation analysis (CANCOR), which is described in the next section (see Gittins 1985 for a more detailed account of this relationship). This interpretation highlights the distinction between PCA and CA; whereas PCA axes maximize linear variation in p-dimensional variable space, CA ordination axes are derived to maximize the correspondence between variables and sampling units. High inertia indicates a strong correspondence between specific combinations of variables and sampling units. CA produces an ordination biplot of variables and sampling units (Figure 5), allowing a visual interpretation of their codependency (Jeffers 1988). Different CA programs scale the variable and individual scores differently, but this does not affect the interpretation of the results (Legendre and Legendre 1998).

CA is currently the most widely used ordination method in ecology and related disciplines (Digby and Kempton 1987; Legendre and Legendre 1998). It is well suited to the summarization of biotic data, which are characteristically nonlinear and contain a large proportion of zeros. However, CA is quite sensitive to outliers and will often highlight unique variable–individual combinations at the expense of summarizing overall data trends (ter Braak 1995, page 109). Simulation tests using artificial data have revealed an additional fault; when a single trend underlies the data, the second CA axis is a simple quadratic function of the first. This

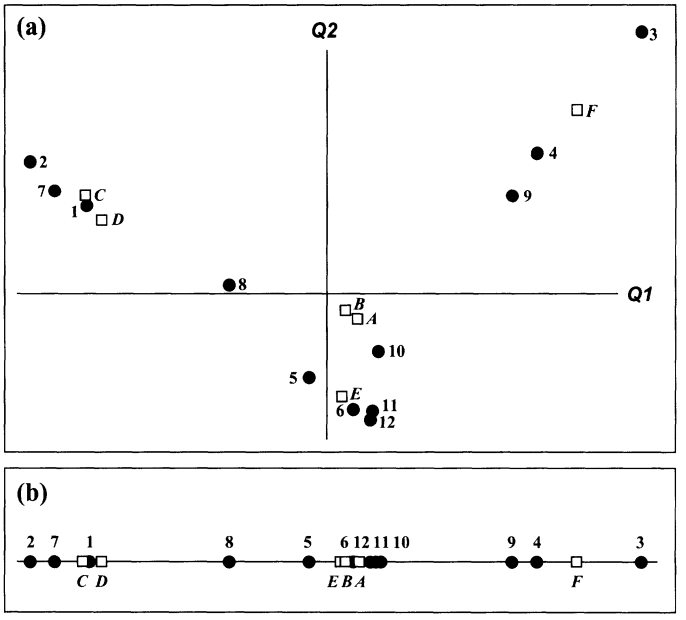

FIGURE 5. Correspondence analysis of twelve sampling units (1 to 12), using data from six species *A–F* in Figure 2. (a) Two-dimensional ordination biplot (*Q1* vs. *Q2*), showing the configuration of sampling units, species, and interrelationship between sampling units and species (e.g., species *F* is most closely associated with sampling units 3, 4, and 9). Total inertia of the data set = 1.121. The first two eigenvalues $\lambda_1 = 0.489$ and $\lambda_2 = 0.373$ together account for 77% of the total inertia, indicating that a two-dimensional ordination summarizes most of the contingency information in the data. (b) A one-dimensional (*Q1*) ordination biplot, which summarizes 43.7% of the total interia.

so-called arch effect is "a mathematical artifact, corresponding to no real structure in the data" (Hill and Gauch 1980). Detrended correspondence analysis (DCA) was explicitly developed to overcome the arch effect (Hill and Gauch 1980). Although widely used, we cannot recommend DCA for general use (see also Legendre and Legendre 1998, page 471). Digby and Kempton (1987) noted that DCA "seems to us to be rather arbitrary and the precise details of the method are hidden in a computer program." Numerous simulation studies have indicated that detrending may distort meaningful CA results (e.g., Jackson and Somers 1991; Kenkel and Orlóci 1986). Detrending should never be applied automatically and should only be used if a demonstrable CA arch effect is observed (Legendre and Legendre 1998).

CA is often performed using the proprietary program CANOCO (ter Braak 1987), which uses Hill's (1973) computationally efficient reciprocal-averaging algorithm and thus avoids eigenanalysis (Digby and Kempton 1987; ter Braak 1995). A recent study indicates some problems with the implementation of CA and DCA algorithms in older versions of CANOCO and offers specific solutions to resolve these problems (Oksanen and Minchin 1997). Conn and Delapp (1983) provided an application of CA to weed science research.

Nonmetric Multidimensional Scaling (NMDS)

In NMDS a map of n individuals in a low-dimensional ordination space is obtained such that the pairwise distances among individuals match as closely as possible to the pairwise distances as measured in the original p-dimensional variable space. Unlike PCoA, the matching of distances in var-

iable and ordination spaces considers only the rankings of the distance values (Digby and Kempton 1987). NMDS does not use eigenanalysis and requires that users specify the dimensionality of the ordination solution before analysis. A computationally intensive algorithm of successive approximations is used to iteratively improve the rank–order relationship between ordination distances and original distances. At each iteration a stress coefficient measuring the correspondence between ranked ordination and variable space distances is computed. Iterations continue until no further reduction in stress is forthcoming. The final result is an optimized mapping of the n individuals in a low-dimensional ordination space. NMDS ordination axes serve only as a relative coordinate system, so unlike PCA, PCoA, and CA ordination axes, they are not interpretable in terms of their importance in summarizing variation or redundancy.

A major advantage of NMDS is its considerable flexibility. As a rank-order–mapping method, it makes no assumptions about the underlying data structure. A wide variety of distance measures are, therefore, acceptable as input into NMDS, including nonlinear and ordinal distance measures (Bradfield and Kenkel 1987; Digby and Kempton 1987). As with PCoA, the lack of variable weights and biplot scores complicates the interpretation of NMDS ordination results. A recent application of NMDS in weed science research is found in Leeson et al. (1999).

Selecting an Ordination Method

Selection of an appropriate ordination method is dependent on both the study objectives and the underlying data structure (Jeffers 1988; Kenkel and Orlóci 1986). Exploratory data analysis will help elucidate the underlying data structure and so guide users in making appropriate decisions regarding which ordination method to use. As a general rule, biotic data are often nonlinear and thus best modeled using CA (or PcoA or NMDS provided specification of an appropriate nonlinear distance measure). By contrast, abiotic data are typically linear (or are easily transformed to meet the linearity assumption) and thus best analyzed using PCA.

A straightforward interpretation of results is paramount when undertaking descriptive modeling of multivariate data, giving ordination methods that produce biplots (PCA, CA) a clear advantage over those that do not (PCoA, NMDS). However, NMDS and PCoA are powerful methods used for analyzing nonstandard data sets (e.g., a matrix of niche overlap values or rank-order data) that cannot be accommodated by PCA and CA. As there is no universal panacea in ordination analysis, it is imperative that users understand the basics of available methods and appreciate both their advantages and limitations.

Predictive Multivariate Modeling: Canonical Methods

The term "canonical" refers to the simplest, most comprehensive form to which the relationship between two variable sets can be reduced without loss of generalization (Legendre and Legendre 1998). Two predictive canonical modeling objectives are distinguished: group discrimination and the analysis of variable set correspondence. CDA is the appropriate statistical model when testing for group discrimination. Consider n sampling units partitioned into g nonoverlapping groups, where p variables are measured on each sampling unit (Figure 3b). CDA tests whether the g groups are statistically different from one another and determines the contribution of each variable to group discrimination. Discriminant analysis is thus closely related to multivariate analysis of variance (ANOVA) (Morrison 1990), the multivariable generalization of univariate ANOVA. Discriminant functions can also be used to optimally assign individuals to one of the several classes (Morrison 1990, Chapter 6).

Other canonical models are used to determine the correspondence between two sets of variables measured on each of the n sampling units (Figure 3c). These include CANCOR, redundancy analysis (RDA), and canonical correspondence analysis (CCA). Canonical correlation is the multivariate extension of product–moment correlation, whereas RDA and CCA are related to multiple regression analysis (ter Braak 1995). RDA is the canonical form of PCA, whereas CCA is the canonical form of CA (Legendre and Legendre 1998).

Canonical Discriminant Analysis

CDA, also known as multiple discriminant analysis or canonical variates analysis, is used in predictive modeling and formal hypothesis testing (Legendre and Legendre 1998). In CDA it is assumed that each of the n sampling units has been assigned to one of the g groups (Gittins 1985). One objective of discriminant analysis, in common with multivariate ANOVA (Morrison 1990), is to determine whether the g groups are statistically different from one another based on the p measured variables. If statistical significance is achieved, discriminant analysis proceeds to find discriminant axes (linear composites, similar to PCA axes) that maximally separate the groups (Figure 6). In addition, discriminant weights are assigned to the p variables on each discriminant axis. Discriminant weights are interpretable as measures of the relative discriminating power of variables in separating groups.

In univariate ANOVA, differences among treatment (group) means are determined using a statistical test based on the ratio of the among- to within-group variances. CDA is essentially the multivariate extension of this concept. Specifically, group discrimination is assessed by comparing the among-groups cross-product matrix **G** and the pooled within-groups cross-product matrix **W.** Numerically, this is achieved through eigenanalysis of the matrix product $\mathbf{W}^{-1}\mathbf{G}$:

$$[\mathbf{W}^{-1}\mathbf{G} - \boldsymbol{\lambda}\mathbf{I}]\mathbf{B} = 0 \qquad [6]$$

where λ is a row vector of t eigenvalues (t is the lesser of $g - 1$ and p), and **B** is an eigenvector matrix containing the discriminant weights. The close relationship between CDA and PCA is apparent upon comparing Equations 1 and 6. The difference is that PCA axes maximize dispersion over the full data set, whereas CDA axes are constrained so as to maximize the dispersion among groups.

In CDA an eigenvalue λ_i is the ratio of the among-groups to the within-groups sum of squares on the discriminant axis i. The discriminant weights for axis i quantify the relative importance of each variable in discriminating among the groups on that axis. Different computer programs scale these weights differently, but their relative values are unaffected. Discriminant weights are most commonly scaled to

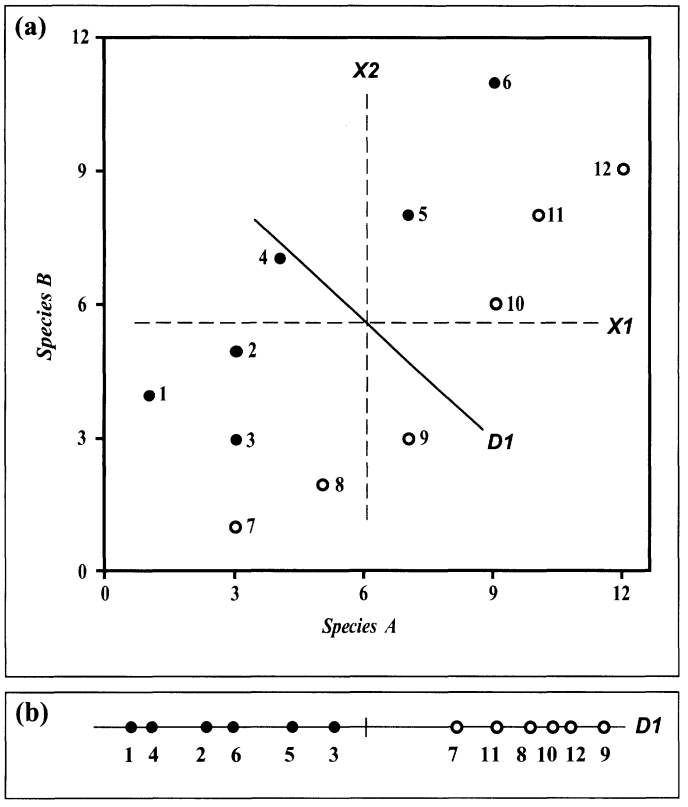

FIGURE 6. Canonical discriminant analysis of two groups (sampling units 1 to 6 and 7 to 12), using data from species *A* and *B* in Figure 2. (a) Scattergram of the 12 sampling units, showing the two groups (open and filled circles) in two-dimensional species space. The original species axes centered on their respective means are shown (dashed lines *X1* and *X2*) together with the single discriminant axis *D1*. Note that *D1* does not correspond to the main variance trend in the data (compare principal component axis *Q1* in Figure 4) because it is constrained to maximally discriminate between the two groups. The eigenvalue (ratio of among-groups to within-groups sum of squares) is $\lambda_1 = 6.77$, which corresponds to $\chi^2 = 18.46$ (df = 2, $P < 0.0001$), indicating significant two-group discrimination. Eigenvector elements are $b_1 = -1.023$ (species *A*) and $b_2 = 1.008$ (species B), indicating that the two species contribute about equally to group discrimination. (b) Projection (discriminant scores) of sampling units 1 to 12 on discriminant axis *D1*.

unit variance within groups (Podani 1994). Like PCA, discriminant scores of sampling units are obtained as linear combinations of the p variables (refer to Equation 2), but with re-scaled discriminant weights as the multiplying factors.

Several assumptions must be met when using CDA for predictive modeling and hypothesis testing:

1. The joint distribution of variables must be approximately multivariate normal (i.e., reasonably symmetric and not too "long tailed"). Outliers strongly affect discriminant analysis results and may lead to erroneous conclusions.
2. The within-group covariance matrices must be homogeneous because they are pooled before eigenanalysis (this is equivalent to the assumption of variance equality in univariate ANOVA). Aberrant covariance matrices will adversely affect the discriminant axis orientation and will produce misleading results.
3. As in PCA, discriminant analysis assumes an underlying linear data structure. Nonlinear data structures will result in highly distorted results.
4. The total number of sampling units must exceed the number of variables measured.

If these assumptions are met, various formal statistical tests are available to determine whether the groups are statistically different and to compute the number of statistically significant discriminant axes (Legendre and Legendre 1998; Morrison 1990). For summary purposes, the sampling units are normally displayed in discriminant axis space. Statistical confidence ellipses for each of the g groups, and biplot scores for the p variables, are often presented as well (Podani 1994).

CDA is a powerful method for exploring and formally testing the statistical significance of various treatment combinations in weed agroecosystem experiments (e.g., Bàrberi et al. 1997, 1998; Derksen et al. 1993). Basic discriminant analysis assumes a complete randomized design. Multivariate designs involving factorial treatment combinations, blocking, and repeated measures can be accommodated in the multivariate ANOVA (Morrison 1990, Chapter 5), but such designs present challenges when applying discriminant analysis. Two-way designs and repeated measures can be accommodated in the discriminant analysis by treating each treatment combination or time interval as a separate discriminant group (e.g., Derksen et al. 1998), but this approach has the potential drawback of ignoring treatment interactions or temporal autocorrelation in the data. Multivariate data from complex experimental designs can also be effectively summarized using the canonical methods described in the following section (e.g., Thomas and Frick 1993).

Canonical Correlation Analysis (CANCOR)

CANCOR determines the linear relationship between two sets of variables, the $\boldsymbol{X}$ set containing p variables and the $\boldsymbol{Y}$ set containing q variables (in developing the method, we will assume that $p \geq q$). The objective of CANCOR is to summarize the correlation between the two variable sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ across the n sampling units. This is accomplished by maximizing the product–moment correlation between a pair of derived linear composites or canonical variates, $\boldsymbol{U}_1$ for variable set $\boldsymbol{X}$ and $\boldsymbol{V}_1$ for variable set $\boldsymbol{Y}$. For example, the scores of individual X_j on the first pair of canonical variates $\boldsymbol{U}_1$ and $\boldsymbol{V}_1$ are given by:

$$U_{1j} = a_{11}(X_{1j} - \bar{X}_1) + a_{12}(X_{2j} - \bar{X}_2) + \cdots + a_{1p}(X_{pj} - \bar{X}_p) \quad \text{and} \qquad [7]$$

$$V_{1j} = b_{11}(Y_{1j} - \bar{Y}_1) + b_{12}(Y_{2j} - \bar{Y}_2) + \cdots + b_{1q}(Y_{qj} - \bar{Y}_q) \qquad [8]$$

The vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ contain canonical weights for the original variables. The method thus derives linear composites $\mathbf{U}_1$ and $\mathbf{V}_1$ such that the correlation between coordinate scores of the n sampling units on $\mathbf{U}_1$ and $\boldsymbol{V}_1$ is maximized (Figure 7). A maximum of t pairs of canonical variates $\mathbf{U}_i$ and $\mathbf{V}_i$ are obtained, where t is the lesser of p and q. The successive canonical variates are obtained subject to their being uncorrelated with previously extracted linear composite pairs.

Canonical correlation involves eigenanalysis of the matrix equation:

$$[\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY} - \boldsymbol{\lambda}\mathbf{I}]\mathbf{B} = 0 \qquad [9]$$

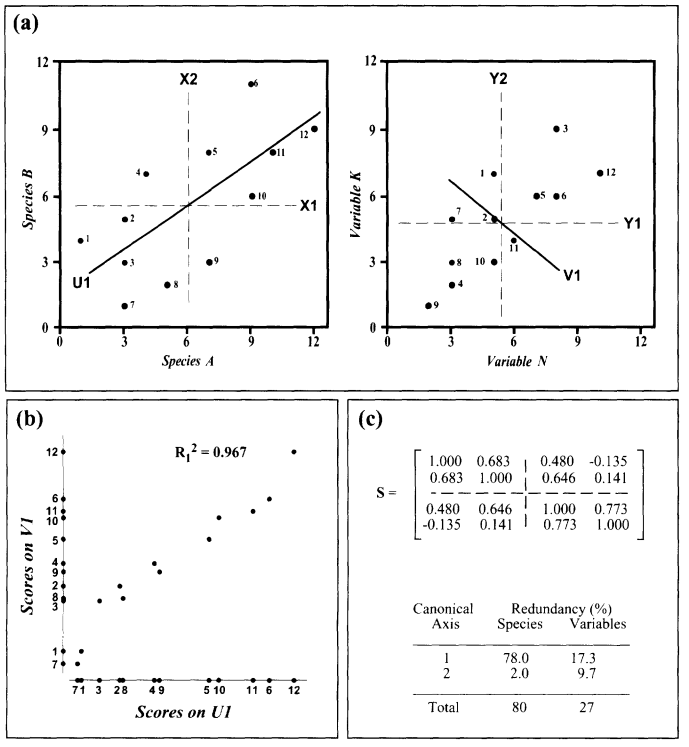

FIGURE 7. Canonical correlation analysis of sampling units 1 to 12, using data from species *A* and *B* and soil variables *N* and *K* in Figure 2. (a) Scattergram of the 12 sampling units in species space (left) and variable space (right), showing the original species and soil variable axes centered on their respective means (dashed lines *X1* and *X2* for species and *Y1* and *Y2* for soil variables). The first pair of canonical axes, *U1* in species space and *V1* in soil variable space, are also shown. (b) Reduced one-dimensional ordinations (scores of the 12 sampling units) on axes *U1* and *V1*. Also shown is a plot of the *U1* vs. *V1* scores, indicating a strong interset correlation ($R_1^2 = 0.967$). The interset correlation for *U2* and *V2* is $R_2^2 = 0.345$. (c) The within and cross-set correlation matrix **S** for the two species and two variables, e.g., the correlations between species *A* and *B* is r = 0.683 and between variables *N* and *K* is r = 0.773. The cross-set correlation between species *A* and variable *N* is r = 0.480. Redundancy values are also summarized. Total redundancy is much higher in species space (80%) than in soil variable space (27%); variation in species space corresponds closely to the canonical relationship between the two data sets (compare *U1* and the principal component *Q1* of the same data, Figure 4), but variation in soil variable space does not.

where $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$ are correlation matrices for variable sets ***X*** and ***Y***, respectively, and $\mathbf{S}_{YX} = \mathbf{S}'_{XY}$ is the matrix of cross-set correlations. The close relationship of CANCOR, CDA, and PCA is apparent upon comparing Equations 1, 6, and 9. The vector $\boldsymbol{\lambda}$ contains the t eigenvalues, which are known as canonical correlations. A canonical correlation $r_{U_iV_i}$ is interpretable as the product–moment correlation between paired canonical variates or alternatively as the multiple correlation between a canonical variate of one variable set and the original variables of the other set.

The eigenvector elements contained in matrix **B** are known as canonical coefficients and are interpretable as canonical weights for the variables of set **X.** Corresponding weights for variable set **Y** are given by:

$$\mathbf{a}_i = [\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\boldsymbol{b}_j]/r_{U_iV_i} \qquad [10]$$

Canonical weights have limitations similar to those of partial coefficients in multiple regression analysis, making them of limited use in interpreting CANCOR results (Gittins 1985). The correlations between observed variables and canonical variates, known as structure correlations, are more stable and easily interpreted (Morrison 1990).

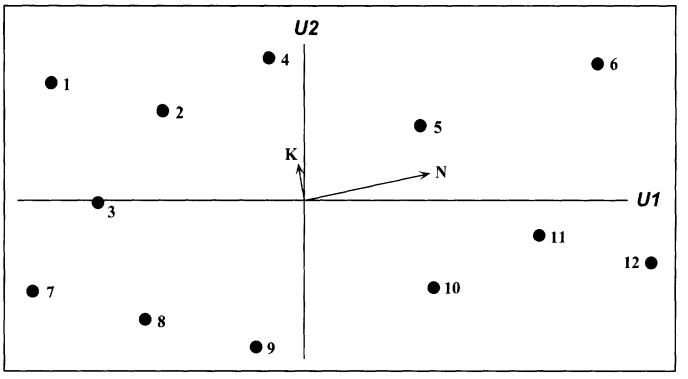

FIGURE 8. Redundancy analysis of sampling units 1 to 12, using data from species *A* and *B* and soil variables *N* and *K* in Figure 2. In this analysis the two species are the response variables. Biplot scores (arrows) of soil variables *N* and *K* are also shown. Axis 1 accounts for 97.6% of the total canonical relationship. The interset correlation for axis 1 is $R_1^2 = 0.967$, and for Axis 2 is $R_2^2 = 0.345$. Note that these are identical to the interset correlations obtained in canonical corrrelation analysis (CANCOR) (Figure 7c). The ratio of the sum of the constrained eigenvalues to the total variance is 80%, which is identical to redundancy of the species data in CANCOR.

Canonical correlation measures the correlation between pairs of derived linear composites, not the original variables themselves. Because the linear composites $\mathbf{U}_i$ and $\mathbf{V}_i$ are not necessarily collinear with the major linear trends (PCA axes) in the **X** and **Y** variable spaces, they may account for only a small proportion of the total variation present in the two data sets. Consider, for example, the case in which a single variable X_i in **X** is highly correlated with a single variable Y_j in **Y,** but all other pairwise correlations among variables in the **X** and **Y** sets are negligible. The canonical correlation will by definition equal or exceed the squared product–moment correlation between X_i and Y_j, even though the remaining correlations are small. This makes canonical correlation a poor measure of the overall relationship between the two variable sets. A direct measure of the interrelatedness of the two variable sets is the redundancy or explained variance (Gittins 1985), which measures the proportion of the total variance of a given variable set that is predictable from the derived canonical variates of the other set. The total redundancy, summed over all canonical variates, measures the proportion of variance in one variable set that is accounted for by the variables of the other set (Figure 7c).

For predictive modeling, tests are available to determine the statistical significance of canonical correlations (Morrison 1990). These tests assume that the data meet the basic assumptions of multivariate linearity and normality. Because canonical correlation summarizes linear relationships, it should only be applied to data with an underlying linear structure. The use of CANCOR in determining relationships between weeds and site characteristics in landscape research has recently been described by Dieleman et al. (2000a, 2000b).

Redundancy Analysis (RDA)

RDA is a canonical or constrained form of PCA (Legendre and Legendre 1998). The objective of RDA is to maximize predictions for a set of response variables **Y** (biotic data), given a set of factor variables **X** (abiotic data). The

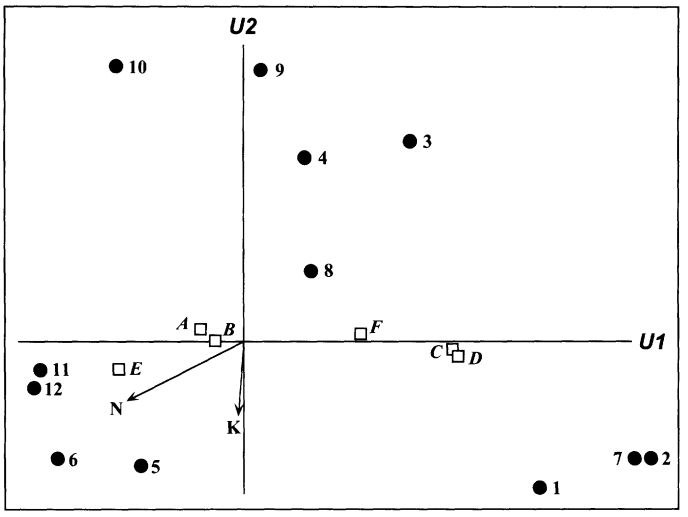

FIGURE 9. Canonical correspondence analysis of sampling units 1 to 12, using data from six species *A–F* and soil variables *N* and *K* in Figure 2. In this analysis the six species are the response variables. Biplot scores (arrows) of soil variables *N* and *K* are also shown. Axis 1 accounts for 89.7% of the total canonical relationship. The interset correlation for Axis 1 is R_1^2 = 0.981 and for Axis 2 is R_2^2 = 0.487. The ratio of the sum of the constrained eigenvalues to the total inertia (0.342/1.121 = 30.5%) is equivalent to the redundancy as used in canonical correlation analysis and redundancy analysis.

method is essentially a PCA in which the sampling unit scores of the response variable set are restricted to be linear combinations of the factor variable set (ter Braak 1995; Wollenberg 1977). The method is, therefore, closely related to multiple regression analysis and produces results similar to CANCOR (Figure 8). An intuitive description of RDA as a multiple regression analysis followed by PCA is given by Legendre and Legendre (1998):

1. Using linear multiple regression, regress each response variable Y_i on the complete set of factor variables **X** and compute fitted multiple regression values.
2. Perform PCA on the set of fitted multiple regression values to obtain a matrix of canonical eigenvectors.
3. Use the canonical eigenvectors to obtain sampling unit scores either in factor space **X** or response space **Y**. Scores in response space are known as weighted averages (WA), whereas those in factor space are known as linear combinations (LC). In most applications, WA scores are more relevant and interpretable (refer to the section on CCA for details).

RDA has not been widely used but has considerable potential in determining the relationship between agronomic treatments and weed community composition (e.g., O'Donovan et al. 1997; Thomas and Frick 1993). Because the method is based on linear multiple regression and PCA, RDA should only be applied to broadly linear data sets (Legendre and Legendre 1998).

Canonical Correspondence Analysis (CCA)

CCA is very similar to RDA, but it is a constrained or canonical form of CA rather than PCA (Legendre and Legendre 1998; ter Braak and Prentice 1988). Like RDA, the method uses multiple regression to select linear combinations of factor variables that best explain variation in ordination scores obtained from the response variables (ter Braak 1995). CCA is widely used in plant ecology to model the canonical relationship between floristic composition (response variables, biotic data set **Y**) and measured environmental variables (factor variables, abiotic data set **X**). Because the method is based on CA, it is well suited to the canonical analysis of nonlinear biotic data sets (Figure 9). As in multiple regression analysis, the inclusion of noisy or trivial factor variables in CCA can result in misleading interpretations (McCune 1997).

Some controversy exists regarding the implementation and interpretation of CCA within the proprietary program CANOCO (Oksanen and Minchin 1997). The program contains a detrending option, but it is not recommended (ter Braak 1987). As in RDA, users have a choice between WA and LC scores. Palmer (1993) recommends using LC scores but under most circumstances we feel otherwise. Because LC scores are obtained directly from the fitted multiple regressions in factor space **X**, two sampling units having identical factor variables will necessarily have identical LC scores even if they share no species (response variables) in common. In most circumstances this extreme degree of constraint is misleading, particularly if CCA is used as a predictive tool and when (as is often the case) only a small subset of all possible factor variables is considered. The use of WA scores is, therefore, recommended for most situations. Legendre and Legendre (1998, page 765) provide an example of predictive modeling in which the use of LC scores is justified.

Although not widely used in the agricultural sciences, CCA is a very powerful multivariate method for descriptive and predictive modeling. CCA has great potential as a method for examining the response of a weed community to various agronomic treatments (e.g., Dale et al. 1992; Del la Fuente et al. 1999; McCloskey et al. 1996). A simple example relating meadow vegetation composition to various agricultural practices is presented in ter Braak (1995, Section 5.4). The proprietary program CANOCO includes more sophisticated methods, such as covariate and partial regression canonical analyses, that are beyond the scope of this review (see Legendre and Legendre 1998; ter Braak and Prentice 1988). Leeson et al. (2000) provide a recent application of partial CCA to weed community analysis.

In applying predictive canonical models, we recommend that users first undertake descriptive analyses of the factor and response variable data sets using ordination methods. By applying CA ordination before CCA, for example, a researcher can objectively determine the effect of canonical constraining. In CCA the ratio of the constrained eigenvalue total to the total inertia is equivalent to redundancy as used in CANCOR.

Selecting a Canonical Method

CDA is appropriate when there exists a known grouping of sampling units, and the objective is to determine the extent to which a set of measured variables can distinguish among these groups. Like PCA, CDA is a linear method and should only be applied to broadly linear data structures with an underlying multivariate normal distribution (Gittins 1985). In addition, CDA requires that the number of sampling units exceed the number of variables. Both of these limitations can be overcome using an approach combining

descriptive and predictive modeling, as explained subsequently. An alternative approach to CDA is to utilize within RDA or CCA "dummy" explanatory variables that code for groups such as agronomic treatments (Legendre and Legendre 1998; ter Braak 1995).

When the objective is to determine the correspondence between two variable sets, CANCOR, RDA, and CCA are the methods of choice. CANCOR is a symmetric method that finds the maximum linear correlation between the factor and the response variables. By contrast, RDA and CCA are asymmetric methods that use multiple regression to determine the extent to which the response variables (biotic data) can be explained by the factor variables (abiotic data). Legendre and Legendre (1998) noted that asymmetric canonical models (RDA and CCA) are more appropriate than the symmetric CANCOR model in most biological applications. RDA is a constrained form of PCA and is therefore appropriate when the two variable sets display linear relationships. By contrast, CCA is a constrained form of CA and should be used when the response variables are nonlinear. Under most circumstances, CCA is the more appropriate method for modeling the relationship between nonlinear species abundance data and environmental factors or agronomic treatments.

Combining Descriptive and Predictive Modeling

As previously discussed, multivariate methods such as PCA, CDA, RDA, and CANCOR, can effectively summarize linear variation but are ill suited to the analysis of nonlinear data structures. Another common problem in multivariate surveys and experiments is that the number of variables is often large and may even exceed the number of sampling units. As the number of variables rises, the potential for variable multicollinearity (a strong interaction among three or more independent variables) increases. Furthermore, the statistical power of multivariate statistical tests is severely compromised when the number of sampling units does not greatly exceed the number of variables (Gittins 1985; Morrison 1990). A stepwise analytic approach can be used to circumvent these problems (Green 1993). As a simple example, consider a biotic data set consisting of p variables and n sampling units divided into three groups, where $p > n$ (Figure 10). As the first step, the data are subjected to CA ordination, which implicitly ignores the underlying group structure. This step produces ordination axes that achieve two important goals: the nonlinear data structure is summarized as new, uncorrelated linear composites, and the number of variables is considerably reduced. The new derived data set consists of scores for the n sampling units on $t \ll p$ derived linear variates (ordination axes). This derived $t \times n$ data set is then used as input into CDA. The assumptions of CDA are now met because the data are linear and the number of variates t is small relative to the number of sampling units. This is a powerful and useful approach, its only potential drawback being that the discriminant analysis is now based on the derived ordination scores rather than on the original variables, complicating the interpretation of the results. This and other stepwise analytic approaches are outlined in Green (1993).

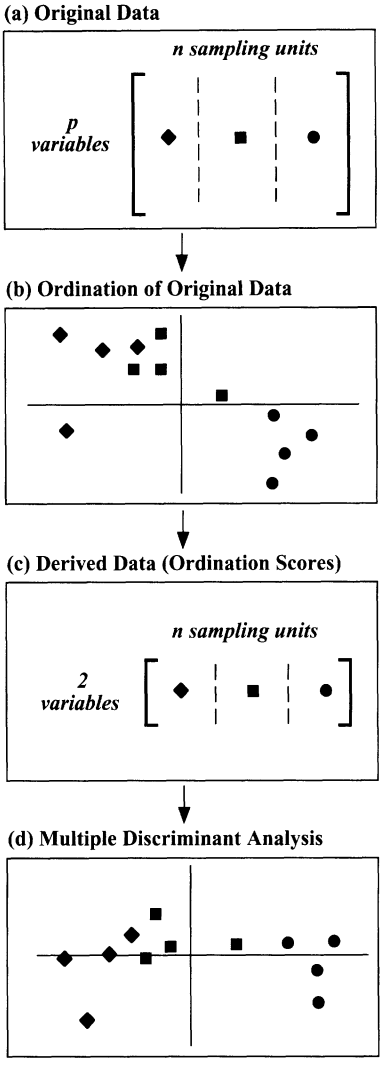

FIGURE 10. A stepwise approach to multivariate analysis. (a) The original data, consisting of cover-abundance measures of p weed species in a series of $n = 12$ plots across three treatments ($n = 4$ for each treatment). The objective is to determine whether the weed communities in the three treatments are statistically different. Unfortunately, the number of species p exceeds the number of plots in each treatment, making it impossible to perform canonical discriminant analysis (CDA) on the original data. Furthermore, the data are highly nonlinear. (b) The original data are subjected to ordination (e.g., principal component analysis or correspondence analysis) to obtain a lower-dimensional representation. (c) Examination of the ordination indicated that a two-dimensional representation captures the essential features of the original data. A derived data set consisting of ordination scores (derived variables) on the first two ordination axes is, therefore, produced. (d) The derived data set is subjected to CDA. Significant discrimination indicates that weed communities in the three treatments are statistically different.

Discussion

Multivariate methods are powerful and sophisticated tools for both descriptive and predictive modeling of complex data structures in weed science. Ordination methods are used primarily to elucidate and summarize underlying trends in many-variable data sets. Discriminant analysis is used to test specific hypotheses regarding the effect of various agronomic treatments on weed community composition, whereas canonical methods are used to examine the relationship between weed community composition and measured environmental factors or agronomic treatment categories (or both). Multivariate methods are, therefore, indis-

pensable to weed science researchers interested in exploring and modeling the structure, composition, and dynamic nature of weed communities.

Acknowledgments

The financial contribution of Agriculture and Agri-Food Canada through the Matching Investment Initiative Program with Dow AgroSciences Canada Inc. is gratefully acknowledged. This research was also supported by an NSERC individual operating grant to N.C.K. We thank two anonymous reviewers for suggestions that improved the clarity of our presentation.

Literature Cited

Bàrberi, P., A. Cozzani, M. Macchia, and E. Bonari. 1998. Size and composition of the weed seedbank under different management systems for continuous maize cropping. Weed Res. 38:319–334.

Bàrberi, P., N. Silvestri, and E. Bonari. 1997. Weed communities of winter wheat as influenced by input level and rotation. Weed Res. 37:301–313.

Benzecri, J.-P. 1992. Correspondence Analysis Handbook. New York: Marcel Dekker. 665 p.

Bradfield, G. E. and N. C. Kenkel. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology 68:750–753.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical Methods for Data Analysis. Belmont, CA: Wadsworth. 395 p.

Cochran, W. G. 1977. Sampling Techniques. 3rd ed. New York: J. Wiley. 428 p.

Conn, J. S. and J. A. Delapp. 1983. Weed species shifts with increasing field age in Alaska. Weed Sci. 31:520–524.

Dale, M.R.T., A. G. Thomas, and E. A. Johns. 1992. Environmental factors including management practices as correlates of weed community composition in spring seeded crops. Can. J. Bot. 70:1931–1939.

Del la Fuente, E. B., S. A. Suárez, C. M. Ghersa, and R.J.C. León. 1999. Soybean weed communities: relationships with cultural history and crop yield. Agron. J. 91:234–241.

Derksen, D. A., G. P. Lafond, A. G. Thomas, H. A. Loeppky, and C. J. Swanton. 1993. Impact of agronomic practices on weed communities: tillage systems. Weed Sci. 41:409–417.

Derksen, D. A., A. G. Thomas, G. P. Lafond, H. A. Loeppky, and C. J. Swanton. 1995. Impact of herbicides on weed community diversity within conservation-tillage systems. Weed Res. 35:311–320.

Derksen, D. A., P. R. Watson, and H. A. Loeppky. 1998. Weed community composition in seedbanks, seedlings and mature plant communities in a multi-year trial in western Canada. Asp. Appl. Biol. 41:43–50.

Dieleman, J. A., D. A. Mortensen, D. D. Buhler, C. A. Cambardella, and T. B. Moorman. 2000a. Identifying associations among site properties and weed species abundance. I. Multivariate analysis. Weed Sci. 48: 567–575.

Dieleman, J. A., D. A. Mortensen, D. D. Buhler, and R. B. Ferguson. 2000b. Identifying associations among site properties and weed species abundance. II. Hypothesis generation. Weed Sci. 48:576–587.

Digby, P.G.N. and R. A. Kempton. 1987. Multivariate Analysis of Ecological Communities. New York: Chapman and Hall. 206 p.

Gabriel, K. R. 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. Pages 147–173 *in* V. Barnett, ed. Interpreting Multivariate Data. Chichester: J. Wiley.

Gittins, R. 1985. Canonical Analysis: A Review with Applications in Ecology. New York: Springer. 351 p.

Goodall, D. W. 1954. Objective methods for the classification of vegetation. II. An essay in the use of factor analysis. Aust. J. Bot. 2:304–324.

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. New York: J. Wiley. 257 p.

Green, R. H. 1993. Relating two sets of variables in environmental studies. Pages 149–163 *in* G. P. Patil and C. R. Rao, eds. Multivariate Environmental Statistics. New York: Elsevier.

Greenacre, M. J. 1984. Theory and Applications of Correspondence Analysis. New York: Academic Press. 364 p.

Hill, M. O. 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61:237–249.

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

Jackson, D. A. 1993. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204–2214.

Jackson, D. A. and K. M. Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. Am. Nat. 137:704–712.

Jeffers, J.N.R. 1978. An Introduction to Systems Analysis: With Ecological Applications. London: University Park Press. 198 p.

Jeffers, J.N.R. 1982. Modeling. New York: Chapman and Hall. 80 p.

Jeffers, J.N.R. 1988. Practitioner's Handbook on the Modeling of Dynamic Change in Ecosystems. New York: J. Wiley. 181 p.

Kenkel, N. C. and L. Orlóci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67:919–928.

Leeson, J. Y., J. W. Sheard, and A. G. Thomas. 1999. Multivariate classification of farming systems for use in integrated pest management studies. Can. J. Plant Sci. 79:647–654.

Leeson, J. Y., J. W. Sheard, and A. G. Thomas. 2000. Weed communities associated with arable Saskatchewan farm management systems. Can. J. Plant Sci. 80:177–185.

Legendre, P. and L. Legendre. 1998. Numerical Ecology. 2nd ed. Amsterdam: Elsevier. 853 p.

Légère, A. and N. Samson. 1999. Relative influence of crop rotation, tillage, and weed management on weed associations in spring barley cropping systems. Weed Sci. 47:112–122.

McCloskey, M., L. G. Firbank, A. R. Watkinson, and D. J. Webb. 1996. The dynamics of experimental arable weed communities under different management practices. J. Veg. Sci. 7:799–808.

McCune, B. 1997. Influence of noisy environmental data on canonical correspondence analysis. Ecology 78:2617–2623.

Morrison, D. F. 1990. Multivariate Statistical Methods. 3rd ed. New York: McGraw-Hill. 495 p.

O'Donovan, J. T., D. W. McAndrew, and A. G. Thomas. 1997. Tillage and nitrogen influence weed population dynamics in barley (*Hordeum vulgare*). Weed Technol. 11:502–509.

Oksanen, J. and P. R. Minchin. 1997. Instability of ordination results under changes in input order: explanations and remedies. J. Veg. Sci. 8:447–454.

Ominski, P. D., M. H. Entz, and N. C. Kenkel. 1999. Weed supression by *Medicago sativa* in subsequent cereal crops: a comparative survey. Weed Sci. 47:282–290.

Orlóci, L. 1966. Goemetric models in ecology. I. The theory and application of some ordination methods. J. Ecol. 54:193–215.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research. The Hague: Junk. 451 p.

Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74:2215–2230.

Podani, J. 1994. Multivariate Data Analysis in Ecology and Systematics. The Hague: SPB. 316 p.

Post, B. J. 1988. Multivariate analysis in weed science. Weed Res. 28:425–430.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya Ser. A 26:329–358.

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. 1995. Ordination. Pages 91–173 *in* R.H.G. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren, eds. Data Analysis in Community and Landscape Ecology. 2nd ed. Cambridge: Cambridge University Press.

ter Braak, C.J.F. and I. C. Prentice. 1988. A theory of gradient analysis. Adv. Ecol. Res. 18:271–317.

Thomas, A. G. and B. L. Frick. 1993. Influence of tillage systems on weed abundance in southwestern Ontario. Weed Technol. 7:699–705.

Tukey, J. W. 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley. 688 p.

Wollenberg, A. L. van den. 1977. Redundancy analysis: an alternative to canonical correlation analysis. Psychometrika 42:207–219.

Zar, J. H. 1974. Biostatistical Analysis. Englewood Cliffs, NJ: Prentice-Hall. 620 p.

Received August 28, 2001, and approved January 9, 2002.