Weed Abundance, Distribution, Diversity, and Community Analyses

Author(s): Roger Nkoa, Micheal D. K. Owen, and Clarence J. Swanton
Source: Weed Science, 63(sp1):64-90.
Published By: Weed Science Society of America
https://doi.org/10.1614/WS-D-13-00075.1
URL: http://www.bioone.org/doi/full/10.1614/WS-D-13-00075.1

Weed Science 2015 Special Issue:64–90

Weed Abundance, Distribution, Diversity, and Community Analyses

Roger Nkoa, Micheal D. K. Owen, and Clarence J. Swanton*

Understanding abundance and distribution of weed species within the landscape of an agroecosystem is an important goal for weed science. Abundance is a measure of the number or frequency of individuals in an area. Distribution is a measure of the geographical range of a weed species. The study of weed population's abundance and distribution is helpful in determining how a population changes over time in response to selective pressures applied by our agronomic practices. Accurate estimates, however, of these two key variables are very important if we are to manage agricultural land both for productivity and for biodiversity.

Biodiversity is generally thought of in terms of number of species and their relative proportion within the plant community. Generally, agronomic practices, including the use of available technologies such as herbicides, limit plant diversity within our cropping systems. Weed species that are able to survive these agronomic selective pressures are ecologically well adapted and invariably become more difficult to manage. In addition, low plant diversity within an agroecosystem can result in the agroecosystem becoming more vulnerable to invasion by new species. Thus, in order to address agronomic or ecological hypotheses regarding weed species abundance and distribution, field experiments must be carefully designed and analyzed in order to reduce the possibility of a Type 1 error. Equally important, however, is an understanding of the limitations of all experimental designs and analyses when developing quantitative data based on field experiments. Limitations such as logistics, time, and funding invariably will influence the procedures for data collection, as well as the geographical scale at which field experiments are conducted.

The scale at which an experiment is to be conducted depends on the question being asked. An experiment to determine the effectiveness of a herbicide on a selected weed species can be conducted at a much smaller scale than an experiment designed to determine if a weed species can be found on specific soil types. How big does an experiment have to be in order to provide useful information? How many samples need to be collected? If a survey is done, how many locations need to be included for the data to be relevant and descriptive? Does the scale of my experiment account for spatial and temporal heterogeneity of the species being studied? Decisions about scale are important; the need to compromise on what would be ideal vs. what can be done realistically must be made. These decisions will influence data analyses, interpretation, and the application of the results. How one plans to analyse data obtained from field work is as critically important as choosing the appropriate experimental design.

In this section we describe and discuss some of the most popular research techniques employed by weed scientists to assess weed abundance, distribution, diversity, and community. We have also introduced some techniques and concepts commonly used in related life science disciplines and in Geographic Information Systems (GIS). These include the concept of "cover," which is popular in general ecology and forestry, and the techniques used in GIS to analyze patterns and distributions of objects or phenomena. The goal here is to strengthen the arsenal of techniques and methods used in weed sciences, and more importantly, to show their applicability in weed sciences through the use of hypothetical example studies. We have provided several landmark references and have drawn realistic illustrations using hypothetical data to enable weed scientists to use GIS techniques and statistics when and where appropriate. Based on the high rate of paper rejections by major scientific journals, this discussion of methods in weed sciences would not be complete without a thorough discussion of the methods of acquisition and analysis of weed data; hence the section on sampling techniques applied to weed sciences. Here, the basic concepts of sample unit and sample size are integrated with sophisticated sampling methods to

DOI: 10.1614/WS-D-13-00075.1

* First and third authors: Research Associate and Professor, Department of Plant Agriculture, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada; second author: Professor, Department of Agronomy, Iowa State University, Ames, IA 50011. Corresponding author's Email: cswanton@uoguelph.ca

match the needs of graduate students and professional researchers. With respect to data analysis, this section attempts to minimize the hurdle of the theoretical aspects of multivariate analysis. Most textbooks and papers on multivariate analysis such as those of Harville (1997), Rao and Rao (1998), and Schott (2005) are written for mathematically advanced readers with academic credits in calculus, matrix algebra, and statistical theory. In this chapter, we bridge the gap between those prerequisites and the level of the average reader by breaking down the matrix algebra concepts that define multivariate techniques. More importantly, the section provides weed science-adapted example studies for illustration, as well as the statistical procedure using the SAS software commands. Out of the many ordination methods available in the literature, we selected two of the most popular in weed science: principal component analysis and canonical discriminant analysis. Other ordination methods (e.g., multidimensional scaling, correspondence analysis) can be found in reference textbooks written by Gauch (1982), and McCune and Grace (2002).

Methods for Assessing Weed Abundance

Abundance measures the quantitative significance of a species in its habitat. It describes the species' success in terms of numbers. Several different methods and techniques of measuring abundance can be used, depending upon the type of species, the habitat (e.g., forest, field), the objectives of the study, and the economic resources of the research team.

Density and Frequency. Density and frequency are the two simplest and most popular methods of measuring abundance. Density (D) measures the number of individuals per unit area, whereas frequency (F) is the proportion of sampling units (e.g., quadrat; field) that contains the species. Thus:

$$D_i = (\Sigma Y_i)/(S_a)$$

and

$$F_i = (\Sigma Z_i)/n$$

where: D_i = density of species i; ΣY_i = number of individual plants of species i contained in the sampling unit (quadrat or field); S_a = Surface area of the sampling unit; F_i = frequency value for species i; ΣZ_i = number of sampling units with species i present; and n = total number of sampling units surveyed.

Issues with Density and Frequency. Recording density is a nondestructive experimental approach. It is very time consuming to get accurate numbers. Once achieved, and unlike frequency, it does provide quantitative information (number of individuals) on the weed species. Recording frequency is fast and nondestructive and is less prone to incorrect estimates than density. In situations where appropriate sampling techniques are employed and sampling points are uniformly distributed across the sampled area, frequency can be a good indicator of the spatial distribution of a species within the sampled area. There are issues associated with using density and frequency, the most prominent of which is the identification of individual plants. Due to the phenomenon of phenotypic plasticity, individuals of the same species can display phenotypic variations and hence appear morphologically different, depending on their developmental or phenological stages and the environment in which they are growing. Thus, measures of density and frequency might exclude individuals that are genotypically similar but somewhat phenotypically different and result in an underestimation of their abundance. It is also possible that in the absence of proper taxonomic information, intraspecific phenotypic variations of weed traits mislead the researcher, resulting in an overestimation of the number of species.

Another issue in using density and frequency as a measure of population abundance is that they do not account for differences in size or weight of the species. Consequently, these measurements might not reflect accurately the ecological importance of a species within a community. For example, at equal density or frequency, larger individuals (such as trees) can have a greater impact on both the community and the environment (e.g., through shading).

Cover and Biomass. Cover is the area of ground covered or the relative proportion of coverage a particular plant species, vegetation layer, or plant form represents when viewed from above. It can be expressed in absolute or relative terms. Absolute cover is the proportion of the ground area, expressed as percent, covered by a particular plant species, vegetation layer, or plant form. It is typically categorized as visual estimates, and as such, can be subjective and relatively inaccurate (Kercher et al.

2003). Common categories are: 0%, 1 to 5%, 5 to 10%, 10 to 25%, 25 to 50%, 50 to 75%, and 75 to 100%. Despite its subjective character, absolute cover is widely used because it provides useful information with relatively low effort. Relative cover is the proportion of the total canopy cover that a particular species, vegetation layer, or plant form represents. For example the statement ''deciduous species represent 75% of the canopy cover'' means deciduous species make up 75% of the canopy (deciduous forest). This method of estimation is more frequent in plant ecology and forestry than in weed biology. As a measure of the proportion of ground occupied by a given weed species or weed community, cover might provide a better indication than density and frequency about the significance of a weed species or a community within a given habitat.

Biomass is the variable used to assess the productivity of a given plant species in a given environment. It is expressed as dry weight per unit area. Biomass is determined by collecting the shoots and/or roots of a species. It is an objective and accurate method; however, it involves destructive sampling and is not convenient for larger organisms such as trees.

Examples of Estimation of Density, Frequency, and Cover. In the following examples (see Figures 1 and 2), the quadrat is the sampling unit and estimates of frequency, density, and cover are shown to be dependent upon the size of the sampling unit and plant distribution.

Example 1: Influence of *quadrat size*. The quadrats are different in size and are distributed randomly within the population (see Figure 1, used with permission from Booth et al. 2010). Frequency is more dependent on quadrat size than other measures of abundance. Use of large quadrats results in more species having 100% frequency, whereas in small quadrats, many frequencies are zero.

Example 2: Influence of plant distribution. Individuals within each population are arranged (a) regularly, (b) in clumps, or (c) randomly. The quadrats are the same size and are randomly distributed within the population (Figure 2, used with permission from Booth et al. 2010). Estimates are different, even though the true values of the three populations are the same. Therefore, under certain circumstances, mean density, frequency, and cover might be of limited value because of sampling bias.

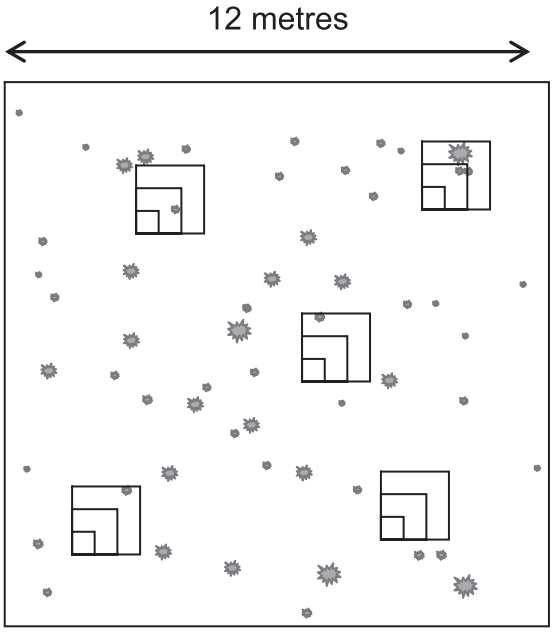
Field Sampling Techniques. Field studies can require sampling across a diverse and large landscape. To save money and time without sacrificing the validity of research findings, a proper sampling method is required. For field sampling, each experimental field is a ''population'' whose mean value is estimated from a sampling distribution of quadrats. The difference between the sampling distribution mean and the field mean constitutes the sampling error. The procedure for selecting the plants to be identified, measured, and used for estimating the field mean is called the field sampling technique. A good field sampling technique is one that produces a small sampling error. To develop a field sampling technique to measure weed abundance and distribution, the research protocol must specify the sampling unit, sample size, and sample design.

Sampling Unit. The sampling unit is the unit from which actual measurements are made. In weed studies, it is usually the quadrat. Important features of an appropriate quadrat are:

- *Stability and definition*. The shape and size of the quadrat must be constant throughout the study, and its boundary with the surroundings should be easily recognized.
- *Ease of measurement*. The measurement of the character of interest or the identification of a weed species should be made easy by the choice of the quadrat.
- *High precision and low cost*. Precision is determined by the reciprocal of the variance of the quadrat estimate. The smaller the variation among quadrat estimates within a given field, the more precise the estimate is, and the size of the quadrat influences the estimate of the character of interest (see ''*Examples of Estimation of Density, Frequency, and Cover*'')

Cost is primarily based on time spent making measurements in the quadrat: the faster the measurement process, the lower the cost. The goal is to maintain a high degree of precision at a reasonable cost, while keeping the variability small among quadrats within a given field.

Sampling Size. The number of quadrats distributed throughout the plot or field is the sampling size. The required number of quadrats for a particular experiment is determined by: (1) The size of the variability among quadrats within the same plot or



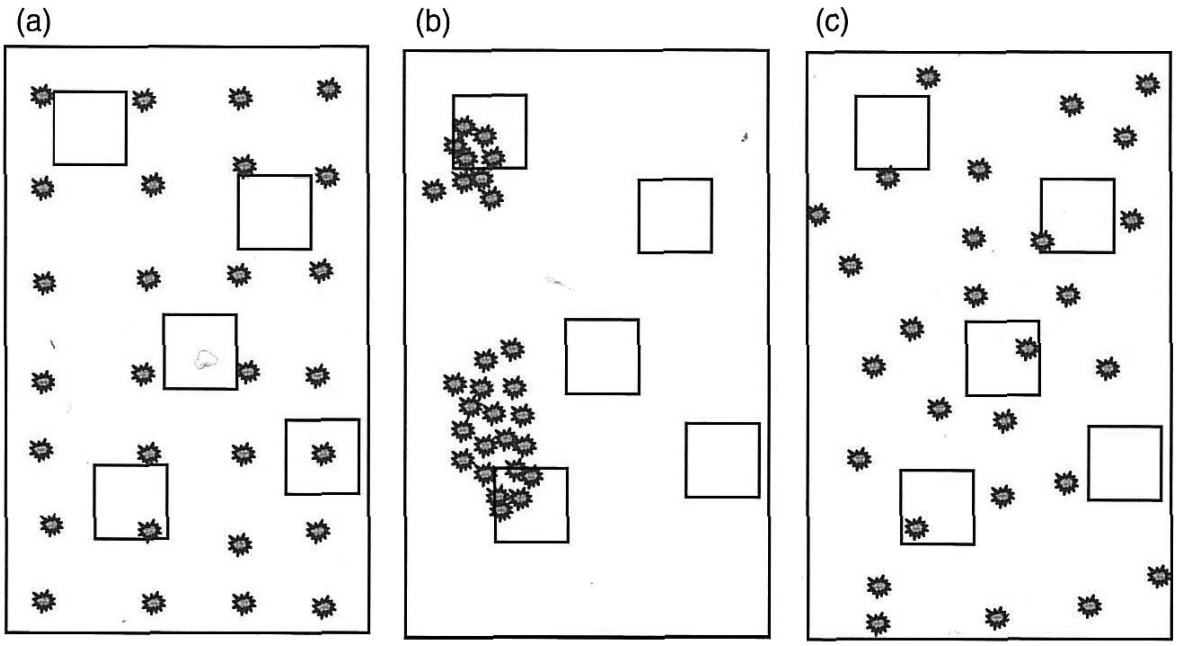
Quadrat Size		Frequency = {no. quadrats with species present / no. quadrats} x 100%	Mean Density = {(Σ no. present per quadrat) / no. quadrats} / quadrat area	Mean Cover = (Σ estimated % cover per quadrat) / no. quadrats
□	0.5m x 0.5m = 0.25m^2	0/5 x 100 = 0%	{(0+0+0+0+0)/5} / 0.25 = 0/m^2	(0+0+0+0+0) / 5 = 0%
□	1.0m x 1.0m = 1m^2	2/5 x 100 = 40%	{(1+1+0+0+0)/5} / 1 = 0.4/m^2	(1.5+.1.5+0+0+0) / 5 = 0.6%
□	1.5m x 1.5m =2.25m^2	4/5 x 100 = 80%	{(1+3+1+1+0)/5} / 2.25 = 0.5/m^2	(1.5+13+1.5+1.5+0) / 5 = 3.5%
True Value:	12m x 12m =144m^2	none	57 / 144 = 0.4/m^2	{(4x.1) + (38x.01) + (15x.005) / 144} x 100% = 0.6%

Figure 1. Estimation of frequency, density, and cover in relation to quadrat size (modified from Booth et al., 2010).

field (sampling variance), and (2) The degree of precision desired for the measurement.

In practice, the amount of variance per sample is generally not known initially. The desired level of precision can, however, be set a priori. The common practice is for the researcher to prescribe the desired level of precision in terms of the margin of error, either of the treatment mean (case of planned replicated experiments) or field/site mean (weed surveys). For example, one might prescribe that the sample estimate should not deviate from the true value by more than 5 or 10%. In the following discussions, we illustrate the procedure for using previously collected data to estimate sample size.

Case Study 1: Estimation of the number of quadrats necessary to estimate weed density in a planned weed management experiment. A weed scientist might wish to evaluate the effect of different weed management strategies on the abundance of weeds. Selection of the proper sampling size (number of quadrats) requires information on the variability of one of the independent variables that best describes weed abundance (density, frequency, cover, and biomass).



Distribution pattern	Frequency	Density	Cover
(a) Regular	40%	0.4/m^2	3.2%
(b) Clumped	40%	2.0/m^2	16.0%
(c) Random	60%	0.6/m^2	4.6%
True value	–	0.7/m^2	5.6%

Figure 2. Estimation of frequency, density, and cover in relation to plant distribution (modified from Booth et al., 2010).

There are three possible sources of data from which the required information can be obtained: data from previous experiments, additional data from ongoing experiments, and data from specifically planned sampling studies. Using data generated from a previous experiment, we demonstrate one method used to calculate the number of quadrats required to meet a set of a priori conditions (see Table 1). Two other methods are described thoroughly in Gomez and Gomez (1984).

The information of primary interest in this case is the treatment mean, that is, the mean weed density (the average weed density across all plots on which the same weed management strategy was applied). Thus, the desired degree of precision will be specified in terms of treatment mean or weed density mean. In such a case, the sample size or the number of quadrats required at α level of significance is computed as shown in Equation 1:

$$n = \frac{\left(Z_{a/2}\right)^2 (v_s)}{r(D^2)(\bar{Y}^2) - \left(Z_{a/2}\right)^2 (v_p)} \quad [1]$$

where n is the required number of quadrats (sample size); Z_α is the value of the standardized normal variate corresponding to the level of significance α (the value Z_α can be obtained from the table of normal curve area); v_s is the sampling variance (i.e., sampling error = variance among quadrats); r is the number of repetitions; D is the prescribed margin of error expressed as a fraction of the treatment mean; $\bar{Y}$ is the mean value of the character of interest; and v_p is the variance between plots of the same treatment (i.e., experimental error).

To illustrate, consider for example an experiment with four replications. Weed densities were recorded on eight different plots on which eight different weed control methods were applied. The researcher wishes to determine the number of quadrats that can achieve an estimate of the treatment mean within 10% of the true value, at 5% significance level. Using data from previous experiments (Table 1) the steps are as follows:

Step 1. Compute the analysis of variance of data from plot sampling based on a randomized complete block design. The results are shown in Table 2.

Table 1. Data on weed density per quadrat obtained from a simple random sample of 12 quadrats per plot in a weed management strategies trial (WMS) involving eight weed control methods and three replications (Rep). (Adapted for illustration purposes only from Gomez and Gomez 1984.)

Weed control methods	Number of weeds per quadrat		
	Rep I	Rep II	Rep III
WMS1	5, 8, 12, 14, 10, 10, 6, 10, 8, 11, 11, 8	10, 13, 10, 13, 11, 11, 12, 5, 10, 7, 14, 5	7, 6, 11, 10, 7, 8, 8, 8, 10, 10, 6, 11
WMS2	11, 11, 11, 12, 4, 12, 8, 14, 8, 7, 9, 9	13, 4, 4, 7, 5, 7, 11, 8, 7, 8, 10, 9	8, 7, 9, 10, 5, 5, 9, 10, 4, 9, 12, 11
WMS3	4, 5, 8, 5, 8, 4, 5, 9, 6, 6, 7, 10	6, 8, 4, 5, 6, 10, 8, 3, 7, 8, 7, 11,	8, 7, 6, 5, 6, 7, 6, 8, 6, 6, 5, 4
WMS4	8, 10, 9, 7, 9, 7, 9, 13, 13, 5, 7, 5	9, 7, 9, 5, 8, 9, 8, 10, 6, 5, 6, 5	8, 10, 7, 6, 7, 6, 9, 8, 6, 4, 5, 7
WMS5	7, 12, 7, 11, 12, 7, 7, 6, 5, 9, 8, 9	9, 7, 6, 8, 4, 8, 8, 9, 8, 9, 6, 7,	9, 3, 4, 6, 5, 3, 9, 7, 9, 6, 6, 7
WMS6	7, 7, 6, 11, 7, 8, 8, 8, 9, 6, 4, 14	8, 10, 7, 6, 8, 8, 10, 5, 7, 5, 8, 7	7, 6, 9, 7, 11, 8, 12, 7, 8, 9, 8, 9,
WMS7	8, 9, 12, 7, 7, 3, 10, 10, 8, 7, 9, 8	8, 6, 7, 8, 9, 9, 14, 8, 9, 11, 6, 7,	10, 4, 8, 9, 4, 6, 7, 4, 3, 4, 4, 6
WMS8	5, 5, 10, 9, 7, 5, 10, 9, 6, 12, 8, 13	8, 8, 8, 3, 13, 13, 7, 12, 9, 9, 8, 11	5, 12, 10, 9, 7, 9, 8, 7, 5, 8, 10, 7

Step 2. Compute the estimates of sampling variance S_1^2 (Equation 2) (i.e., the variance between quadrats within a plot) and of experimental error S_2^2 (Equation 3) (i.e., the variance between plots of the same treatment) as:

$$S_1^2 = MS_1 \quad [2]$$

$$S_2^2 = \frac{MS_2 - MS_1}{n} \quad [3]$$

where MS_1 is the sampling error mean square; MS_2 is the experimental error mean square in the analysis of variance computed in step 1; and n is the sample size, that is, the number of quadrats per plot.

Using the results of the analysis of variance (Table 2), the estimates of sampling variance (S_1^2) and experimental error (S_2^2) are computed as:

$$S_1^2 = 5.0429$$

$$S_2^2 = \frac{7.3993 - 5.0429}{12} = 0.1964$$

If given the average number of weed plants per quadrat $\bar{Y} = 8$; $Z_{\alpha/2} = 1.96$; $v_s = 5.0429$; $vp = 0.1964$; $D = 0.1$; and $r = 4$; the required number of quadrats, at 5% level of significance and 10% margin of error, is then computed from Equation 1 as:

$$n = \frac{(1.96)^2(5.0429)}{4(0.1)^2(8)^2 - (1.96)^2(0.1964)}$$

$$n = 10.7 \approx 11 \text{ quadrats/plot}$$

Case Study 2: Weed surveys. Now, consider that the weed scientist wishes to determine the abundance of weeds within an ecological land unit comprising f selected fields. Equation 1 can be written as shown in Equation 4:

$$n = \frac{\left(Z_{a/2}\right)^2(v_s)}{f(D^2)(\bar{Y}^2) - \left(Z_{a/2}\right)^2\left(v_f\right)} \quad [4]$$

where n is the required number of quadrats (sample size); Z_α is the value of the standardized normal variate corresponding to the level of significance α; v_s is the sampling variance (i.e., sampling error = variance among quadrats); f is the number of selected fields; D is the prescribed margin of error expressed as a fraction of the treatment mean; $\bar{Y}$ is the mean value of the character of interest (e.g., weed density); and v_f is the variance between fields.

In the case where the survey involves a single field, then $f = 1$ and $v_f = 0$; Equation 4 becomes Equation 5:

Table 2. Analysis of variance (randomized complete block design) of data from Table 1. (Source: Gomez and Gomez 1984).

Source of variation	Degree of freedom	Sum of squares	Mean square
Replication	2	53.5208	26.7604
Variety	7	191.0556	27.2937
Experimental error	14	103.5903	7.3993
Sampling error	264	1,331.3333	5.0429

$$n = \frac{\left(Z_{a/2}\right)^2(\sigma)}{D^2\bar{Y}^2} \quad [5]$$

where n is the required number of quadrats (sample size); Z_α is the value of the standardized normal variate corresponding to the level of significance α; σ is the field variance (i.e., population variance); D is the prescribed margin of error expressed as a fraction of the treatment mean; and $\bar{Y}$ is the mean value of the character of interest (e.g., weed density). The value of σ is usually unknown. However, it can also be estimated by the standard deviation, s, from a prior or preliminary sample: $\sigma = s$, when $n \geq 30$. (McClave and Dietrich 1988).

Sampling Designs. A field sampling design specifies the manner in which the n quadrats are to be selected from the whole plot or field. An appropriate sampling design must satisfy the following requirements: (1) the precision of the estimate must match the precision needed to fulfill the experimental goal; and (2) the cost of its implementation must be within the financial, human, and logistic resources available to the researcher.

There are four commonly used sampling designs: simple random sampling, stratified random sampling, systematic sampling, and timed-meander sampling.

Simple Random Sampling. This is a method of selecting n quadrats from each plot or field consisting of a total of N quadrats. The selection of the n quadrats is done in such a way that each of the N quadrats in the plot or field is given the same chance of being drawn. In practice, two of the most commonly used procedures for selecting n quadrats per plot/field are the random-number technique and the random-pair technique.

Random-number technique. This technique is most useful in cropped experiments, when the plot/field can be divided into N distinct quadrats. We illustrate the process used to apply the random-number technique with the case of an experimental plot.

- First, divide the plot into N distinctly differentiable quadrats, and assign a number from 1 to N to each quadrat in the plot. For our example, the plot is divided into $N = 35$ quadrats, each of which is assigned a unique number from 1 to 35 (Figure 3).
- Next, randomly select n distinctly different numbers, either by means of a table of random numbers (Snedecor and Cochran 1967) or by means of a computer program that produces such a table. For our example, $n = 5$ random numbers (each within the range of 1 to 35) are selected from the table of random numbers. The five random numbers selected might be: 15, 6, 28, 17, and 24.
- Finally, use, as the sample, all the quadrats whose assigned numbers (step 1) correspond to the random numbers selected in step 2. For our example (Figure 3), the five quadrats in the plot whose assigned numbers are 15, 6, 28, 17, and 24 are used as the sample.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	**15**
16	**17**	18	19	20
21	22	23	**24**	25
26	27	**28**	29	30
31	32	33	34	35

Figure 3. Plot location of five randomly selected quadrats (Bold number), using the random-number technique.

Random-pair sampling technique. This technique is applicable whether or not the plot/field can be divided into N quadrats, which makes this technique more popular than the random-number method. We illustrate the procedure with the typical case involving weeds where the field cannot be divided into N distinct quadrats. Consider a case in which a sample of ten 1- by 1-m quadrats is to be selected at random from a field measuring 50 by 100 m. The steps involved in applying the random-pair technique to select a random sample of $n = 10$ quadrats are:

- First, specify the width (W) and length (L) of the field, using the same measurement unit as that of the quadrat. For our example, the meter is used as the measurement unit because the quadrat is defined in that scale. Thus, the field width (W) and length (L) are specified as 50 m and 100 m.
- Second, select n (= 10) random pairs of numbers from the table of random numbers, with the first

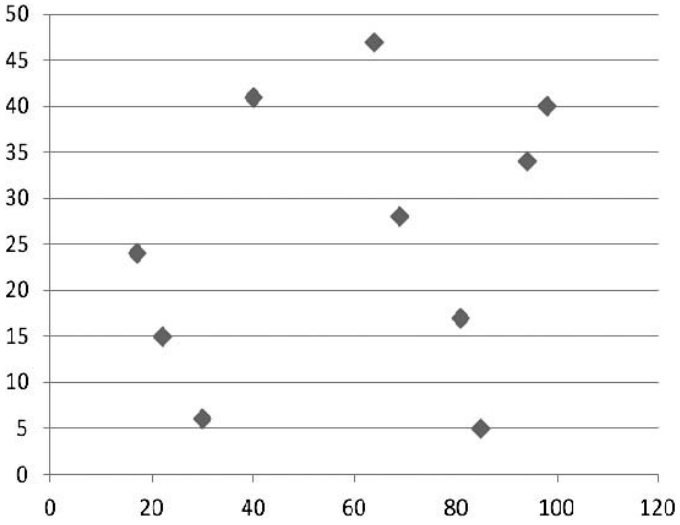


Figure 4. Field location of 10 randomly selected 1- by 1-m-quadrats, using the random-pair technique for a field measuring 50 by 100 m.

number of the pair lying between 1 and W (= 50) and the second number lying between 1 and L (=100). For our example, the 10 random pairs of numbers might be: (15, 22);(5, 85); (41, 40); (28, 69); (17, 81); (40, 98); (34, 94); (6, 30); (47; 64); (24; 17).

- Third, use the point of intersection of each of the random pairs of numbers to represent the center of each selected quadrat. For our example, the first random pair of (15, 22) is the first selected 1- by 1-m quadrat whose center is at the intersection of 15 m along the width of the field and the 22 m along the length of the field (Figure 4). The rest of the selected quadrats can be identified in the similar manner. Points that fall beyond the delineation of the field are discarded and substituted by new ones that fall within the field.

Stratified Random Sampling. In this technique, an ecological land unit or a field population is first divided into relatively homogeneous, nonoverlapping k subpopulations called strata (Figure 5) before a set of m_k sampling units are selected randomly from each stratum. Thus, the total number n of sampling units per field is: $n = m_1 + m_2 + \ldots + m_k$.

The technique is useful where there is a large variation between sampling units and where environmental and anthropogenic factors such as climate change, soil, landscape, cropping systems management induce a consistent pattern of variability across the landscape. In weed science, stratification can be very convenient for sampling. The reasons for this include the following:

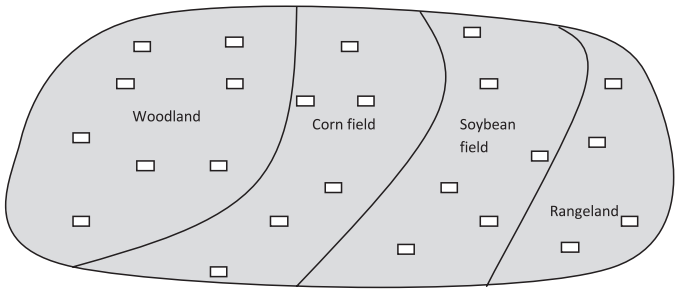


Figure 5. Hypothetical example of stratified random sampling in a farmland comprising four strata: a woodland, a rangeland, a corn field, and soybean field.

- If information of known precision is needed for certain strata of the ecological land unit (for instance, weed richness within a farmland comprising fields cropped with different species and a natural system such as a woodland), then it is advisable to treat each stratum as a ''population'' in its own right.
- When strata are not apparent, it might be more effective to create your own strata boundaries in order to carry out a weed survey, because different teams can work simultaneously on different strata.
- Sampling issues might differ from one stratum to another. For example, sampling in a cropland would offer a different kind of challenge than sampling in a natural system.
- Stratification can improve the precision of the estimates of the population parameters. If variability within each stratum is minimal, that is, if the measurements vary little from one unit to another, then a precise estimate of any stratum mean can be obtained from a small sample in that stratum. A precise estimate of the whole population can then be obtained by combining estimates of individual strata.

Systematic Sampling. In systematic sampling, a field or plot is divided into N units that are numbered 1 to N in some order. To select a sample of n units, a unit is taken at random from the first k units and every *kth* unit thereafter. For instance, if $k = 4$ and if the first unit drawn is number 2, then the subsequent units are numbers 6, 10, 14, and so on (Figure 6). The selection of the first unit determines the whole sample. This type of sample is termed ''every kth systematic sample.'' The apparent advantages of the systematic method over simple random sampling and stratified random sampling are: (1) it is easier to draw a sample and speedier to execute without mistakes; and (2) the fact that a systematic sample is spread more evenly over the

1	**2**	3	4	5	**6**	7	8	9
18	17	16	15	**14**	13	12	11	**10**
19	20	21	**22**	23	24	25	**26**	27
36	35	**34**	33	32	31	**30**	29	28

Figure 6. Example of systematic sampling on a plot for $N = 36$; $k = 4$; first draw = 2.

population has sometimes made it considerably more precise than random sampling and stratified random sampling.

Timed-meander Sampling. This method entails thoroughly following a meandering walk through a delineated field to tally weed species (Figure 7). Every 10 (or fewer) min, species and time are recorded on a field data form as they are encountered, until the number of new species listed equals zero in the last 10 min of walking. If the number of new species listed does not decrease to zero in the last 10 min of walking, additional time is added. The procedure divides the species list into sets of species recorded or collected during each time interval.

This method samples 100% of the field, and thus there are no designated plots per se, and as such, any variable based on surface area cannot be analyzed. However, the timed-meander technique is very effective in studies where it is deemed to be the best approach to maximizing coverage and the potential for identifying rare or invasive weed species (Goff et al. 1982; Huebner 2007; Penskar 1991).

In this section, we have discussed the key elements of field sampling techniques (sampling unit, sample size, sample design). The researcher must bear in mind that there is an abstract component to field sampling that is of importance: the researcher's judgement. For instance, the selection of sampling points through random numbers does not guarantee an even distribution of samples across the field. Common sense should therefore guide the researcher to discard numbers that lead to clumped sampling points. Another example that merits special consideration, especially in the case of floristic studies, is the timing of sampling. Some weed species might not be fully identifiable at the time of sampling. The weed biologist might need to consider resampling each field unit several times during the growing season to account for differential emergence timing and growth and development.

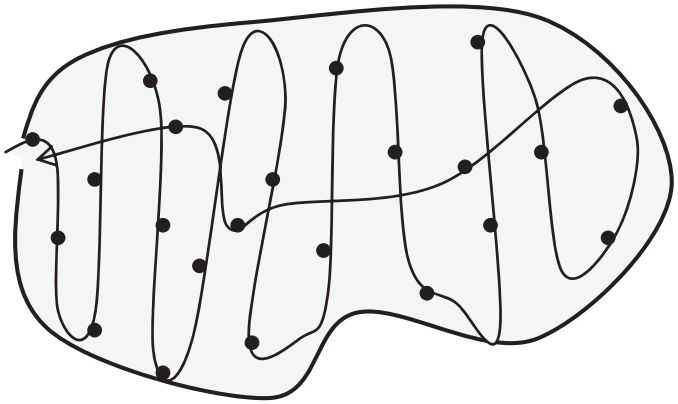


Figure 7. An example of a meander itinerary in a sampled field.

Methods for Estimating and Mapping Weed Distribution

A weed species' distribution is its natural geographic range. It is the description of where the species naturally occurs, or where it has been recorded (Gaston 1991). Because a species might not always occupy all possible sites in which it can survive, it is therefore important to distinguish among three basic concepts: the extent of occurrence, the area of occupancy, and the potential distribution. The extent of occurrence is the entire area that lies within the outer boundaries of a range (Gaston and Fuller 2009). The area of occupancy is that area within the extent of occurrence where the species actually occurs. The potential distribution is the area in which abiotic factors would allow a species to survive. A species' distribution is a dynamic phenomenon; it changes over time as the result of such factors as climatic (e.g., climatic change), anthropogenic (e.g., land use change), or ecological (e.g., succession change, disease outbreaks). Changes in a species' distribution can provide critical information such as: species' expansion or contraction, predictability of occurrence, effectiveness of control measures, habitat preferences, and dispersal mechanisms. The delineation of the spread of a weed species is achieved through data collection and mapping.

Data Collection. Data on the actual distribution of a weed species can be collected directly from the field or indirectly from public records, such as government documents, herbaria, and academic and research publications.

Direct methods involve data collection, analysis, and representation by the researcher conducting the study. Thanks to the technological revolution in the development of GIS, it is now possible to map data,

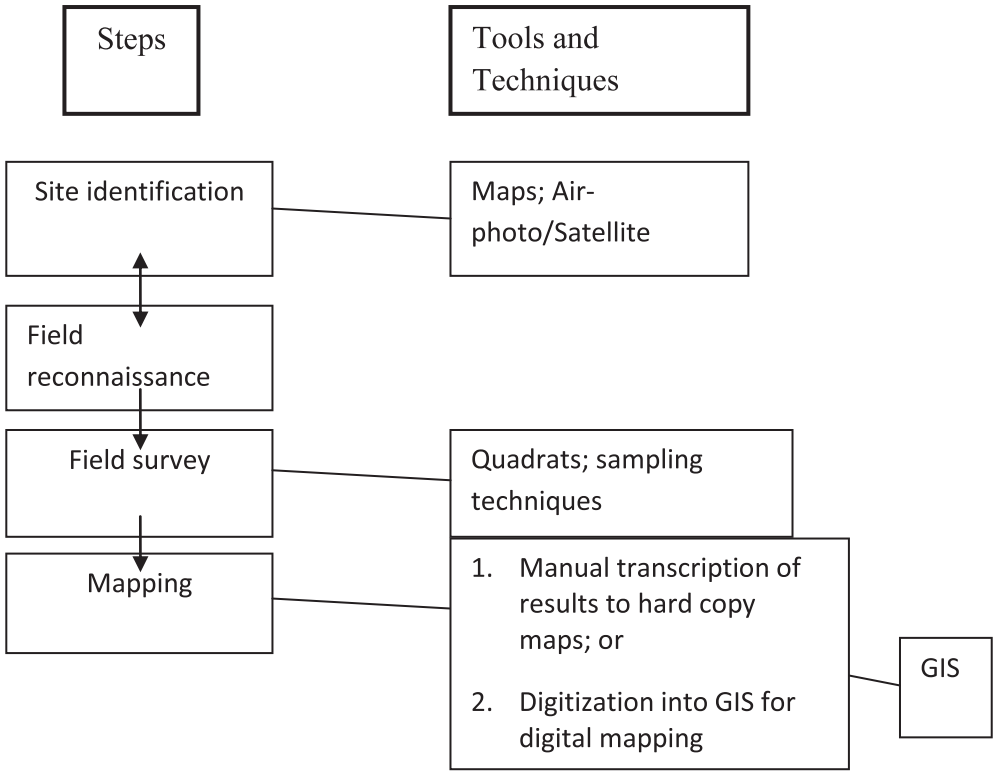


Figure 8. Schematic representation of the process for estimating and mapping weed distribution.

and above all, to spatially and temporally analyze maps with a relative ease (see next section). Four main steps can be distinguished: site identification, field reconnaissance, field survey, and mapping (see Figure 8).

Indirect methods, such as consultation of public records are inexpensive and allow for construction of historical distributions (Forcella and Harvey 1988; Pearman et al. 2008; Thomson et al. 1987). However, this method of data collection can suffer several setbacks, including: (1) few data from public records are digitized, which make them difficult to collect and organize; (2) collection bias resulting from different sources of data with different sampling accuracies and precisions (for instance, some sites or species might be more intensively sampled over time and space than others, making them overrepresented on a map); and (3) public records do not guarantee time continuity and thus, there could be time gaps when no samples were collected, making it difficult to follow the floristic process.

Geographic Information System. GIS is a computer-based system specially designed to manage geospatial data and to use this data to solve spatial problems (Lo and Yeung 2007). The field or landscape upon which a survey is to be conducted can be geographically referenced and then used to create geospatial data based upon information gathered from the survey. This information can then be used to map the abundance and distribution of a specific weed species or population in order to determine dispersal routes of an invasive species, or the effectiveness of specific weed control measures to limit the invasiveness of a particular weed species. In the past, geospatial data were not easily accessible. Today, thanks to advances in computer sciences and the internet, GIS users can take advantage of the huge number of databases in existence from different organizations. Specifically, the researcher has several options for acquiring or digitally converting spatial data for use in computer systems: digitizing existing maps; purchasing digitized data from government agencies or commercial data suppliers; or collecting new data using GPS-based attribute data loggers and photogrammetric and remote sensing methods.

In-House Digitizing. Once the survey data has transcribed onto a hard-copy map, the map can be converted into digital format using data-capture devices, such as a digitizer or a scanner. Map digitizing is the conventional method; although its importance has diminished considerably in recent years. This method, however, is still the most

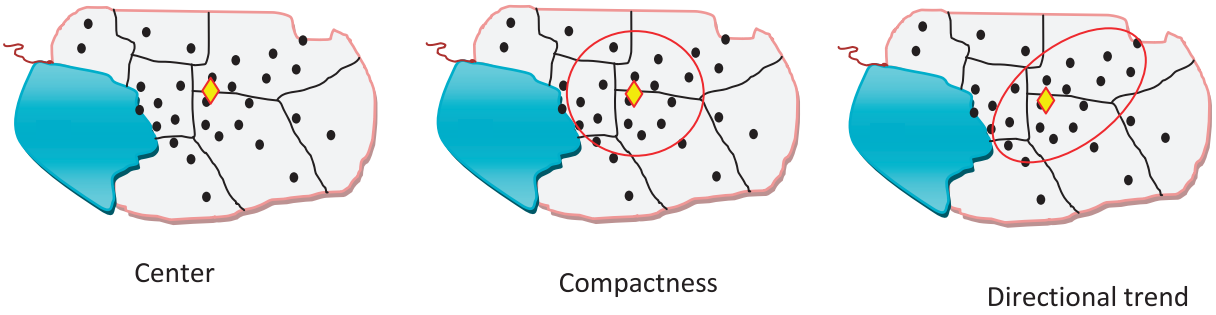


Figure 9. Illustration of weed distribution characteristics (center, compactness, and directional trend) of an invasive weed species in a surveyed area.

convenient way to create a geospatial database "in-house." It is particularly suitable for GIS applications that require project-specific data (i.e., data unlikely to be available from external sources).

Scanning, on the other hand, is often referred to as "screen digitizing" or "heads-up digitizing" to distinguish it from conventional table digitizing. Significant advances in hardware and software technologies and reduced hardware costs have made scanning readily accessible and affordable to computer users. Scanning is now the most widely used technology for converting hard-copy maps into digital formats (Flanagan et al. 1994; Mayo 1994). Both map-digitizing and scanning procedures are thoroughly described by Lo and Yeung (2007).

GIS Implementation in Weed Science. Most weed science studies have used GIS to geographically locate specific weed species and populations of interest. GIS can also be used for analyses to address questions such as why weeds are where they are and how their distribution and abundance are related to climatic, environmental, or anthropogenic factors. Resources needed to implement such analyses using GIS can be found at the Environmental Systems Research Institute (ESRI): geodatabases, softwares (ArcGIS), online GIS courses, GIS guide books (Mitchell 1999; 2009), and GIS tutorials books (Allen 2009; Gorr and Kurland 2008). Specifically, tools in ArcGIS can help weed scientist's measure geographic distributions, identify patterns and clusters, and analyze geographic relationships.

Measuring Geographic Distributions. Measuring weed distribution using GIS allows the weed scientist to calculate and display characteristics of the weed distribution such as its center (i.e., the average $\times$ coordinate and average y coordinate for the sampling point in the study area), compactness (clustering or dispersion around the center), or orientation (trend in a particular direction) (Figure 9). These distribution characteristics can shed light, for example, on the point of entrance of an invasive species, and the most suitable abiotic and biotic environment for growth and development, as well as possible dispersal routes.

Identifying Patterns. Identifying distribution patterns of specific weed species or populations can provide insights into the habitat requirements and enable the researcher to compare and track changes over time. For instance, weed species A can be found to be dispersed along the coastal line, suggesting that its preferential habitat is determined by the sea environment; however, species B forms clusters inland, suggesting that this species avoids the sea environment and prefers a drier range of habitats (Figure 10). GIS statistics such as the *k*-statistic, or nearest-neighbour index (see Mitchell 1999; 2009) can be used to measure and display patterns. These statistics can be used to compare the actual weed distribution (observed distribution) to a hypothetical random distribution of the same number of sampling points over the same area. The extent to which the observed distribution deviates from the random distribution is the extent to which the pattern is more clustered or more dispersed than the random distribution. The validity of the analysis

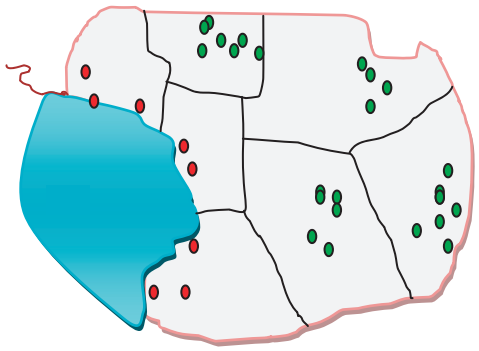


Figure 10. The patterns of weed species A (red) dispersed along the costal line, and species B (green) clustered inland.

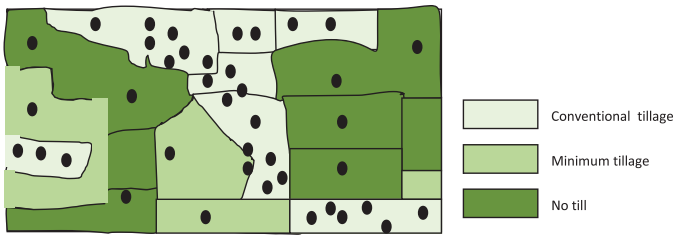


Figure 11. Hypothetical relationship between tillage method and weed density.

is expressed by the probability that a pattern isn't simply due to chance.

Analyzing Geographic Relationships. Using GIS, one can also analyze relationships describing weed presence/absence, distribution, abundance, frequency, and with abiotic and biotic variables in order to better understand spatial relationships at the landscape level. Thus, for example, GIS can use statistics such as the Pearson's or the Spearman's rank correlation coefficient to measure, test, and map the spatial relationship between weed distribution and tillage systems. In this hypothetical case (Figure 11), a spatial analysis can show that weeds are more abundant in areas where conventional tillage is practised as opposed to areas where conservation tillage has been adopted.

Methods for the Evaluation of Diversity

Diversity can be explored at several different scales from number of species per unit area to genetic diversity. For this discussion, we focus on species diversity. Species diversity is described by two components: richness and evenness. Richness is the number of species present in an area or in a community, whereas evenness specifies the abundance of each species in a community. Evenness provides information on whether a community is dominated by one or more species or whether the species within the community are represented by approximately equal numbers (Booth et al. 2010).

Measurement of Diversity. Several methods of measuring diversity, with varying advantages and disadvantages, can be found in the literature (Conroy and Noon 1996; Cousins 1991; Magurran 1988; Schlesinger et al. 1994; Stiling 1999; Wilson et al. 1999; Yorks and Dabydeen 1998). Diversity can be estimated within a given community (alpha-diversity) or between communities (beta-diversity).

Within-Community Diversity. Three indices are commonly used to estimate the within-community-diversity: the Margalef's Diversity Index (D_{Mg}), the Shannon–Wiener Diversity Index (H'), and the Simpson's Dominance Index (D). The Margalef's Diversity Index is a quick method of estimating the species diversity based on richness. It is sensitive to the sampling technique (sampling unit, sampling size, and sampling design). The Shannon–Wiener Diversity Index, unlike the Margalef's index, specifies both species richness and evenness (Magnussen and Boyle 1995; Magurran 1988). This method is moderately sensitive to sample size. Lastly, the Simpson's Dominance Index measures the state of dominance within the community. This method is less sensitive to sample size; however, it does not provide an assessment of species richness, but is useful when describing evenness.

In the following discussion, we show how to calculate three within-community diversity indices: the Margalef's Diversity Index (D_{Mg}), the Shannon–Wiener Diversity Index (H'), and the Simpson's Dominance Index (D), for two communities located within the same meadow but with a different soil type (a cultural mineral soil habitat and a cultural high organic soil habitat within the same meadow), both dominated by invasive species. We use the following symbols:

- n = population density or number,
- n_i = density or number of the ith species,
- N = total number of individuals of all species in the community,
- S = species richness (i.e., total number of species),
- Σ means: sum of all the factors that follow, and
- p_i = proportional abundance or relative frequency of the ith species,

The set of data given in Table 3 is used for all the calculations:

Margalef's Diversity Index (D_{Mg}). The formula (Equation 6) and calculations follow.

$$D_{Mg} = (S-1)/\ln(N) \quad [6]$$

$$\begin{aligned} D_{Mg\,(\text{mineral meadow})} &= (S-1)/\ln(N) \\ &= (9-1)/\ln(278) \\ &= 8/5.628 \\ &= 1.421 \end{aligned}$$

Table 3. Hypothetical weed populations identified within a meadow consisting of two distinct habitats: cultural mineral soil meadow and cultural high organic soil meadow. Data were collected using a systematic random sampling method. (Source: Fictitious data, for illustration purposes only).

Weed species	Common name	Mineral meadow n_i	Organic meadow n_i
Polygonum convolvulus L.	Wild buckwheat	53	0
Amaranthus retroflexus L.	Redroot pigweed	11	41
Chenopodium album L.	Common lambsquarters	15	58
Brassica kaber (DC.) L. C. Wheeler (*Sinapis arvensis* L.)	Wild mustard	16	48
Taraxacum officinale G. H. Weber ex Wiggers	Common dandelion	36	59
Stellaria media (L.) Vill.	Common chickweed	26	78
Agropyron repens (L.) Beauv. [*Elymus repens* (L.) Gould]	Quack grass	40	0
Medicago sativa L.	Alfalfa	27	32
Avena fatua L.	Wild oat	54	0
N^a		278	316
S^a		9	6

[a] Abbreviations: *N*, total number of all species in the community; *S*, species richness (total number of species).

$$D_{Mg\,(\text{organic meadow})} = (S-1)/\ln(N)$$
$$= (6-1)/\ln(316)$$
$$= 5/5.756$$
$$= 0.869$$

Margalef's Diversity Index calculations indicate that the mineral meadow is more diverse than its counterpart organic meadow. This can be intuitively obvious when the size of the data is not large.

The Shannon–Wiener Diversity Index (H′). The formulas for H' (Equation 7) and evenness (E, Equation 8), and calculations follow:

$$H' = \Sigma[-p_i(\ln p_i)] \quad [7]$$

The values of H' are calculated by summing all the $-p_i(ln\ p_i)$ values for each meadow community (Table 4). Then, using the calculated values of H', calculate evenness (E):

$$E = H'/\ln S \quad [8]$$

Evenness (E) can now be calculated using the calculated values of H' from Table 4:

$$E = H'/\ln S$$
$$E_{(\text{mineral meadow})} = H'_{(\text{mineral meadow})}/\ln S_{(\text{mineral meadow})}$$
$$= 2.075/\ln 9 = 2.075/2.197 = 0.944$$
$$E_{(\text{organic meadow})} = H'_{(\text{organic meadow})}/\ln S_{(\text{organic meadow})}$$
$$= 1.753/\ln 6 = 1.753/1.792 = 0.978$$

Overall, these calculations indicate that $H'_{(\text{mineral meadow})} = 2.075$, and $H'_{(\text{organic meadow})} = 1.753$; and that $E_{(\text{mineral meadow})} = 0.944$, and $E_{(\text{organic meadow})} = 0.978$.

Table 4. A hypothetical example of how to calculate the Shannon-Wiener Diversity Index (H′). This example compares the weed species found in two portions of the same habitat: the cultural mineral soil meadow and the cultural high organic soil meadow. (Source: Fictitious data, for illustration purposes only).

	Mineral meadow[a]				Organic meadow[a]			
Weed species	n_i	p_i	lnp_i	$-p_i(lnp_i)$	n_i	p_i	lnp_i	$-p(lnp_i)$
Polygonum convolvulus L.	53	0.191	−1.655	0.316	0	0.000		
Amaranthus retroflexus L.	11	0.040	−3.219	0.129	41	0.130	−2.042	0.265
Chenopodium album L.	15	0.054	−2.919	0.158	58	0.184	−1.695	0.311
Brassica kaber (DC.) L. C. Wheeler	16	0.058	−2.847	0.165	48	0.152	−1.885	0.286
Taraxacum officinale G. H. Webber ex Wiggers	36	0.129	−2.048	0.264	59	0.187	−1.678	0.313
Agropyron repens (L.) Beauv.	40	0.144	−1.938	0.279	0	0.000		
Stellaria media (L.) Vill.	26	0.093	−2.375	0.220	78	0.247	−1.399	0.345
Medicago sativa L.	27	0.097	−2.333	0.226	32	0.101	−2.291	0.232
Avena fatua L.	54	0.194	−1.640	0.318	0	0.000		
Σ (sum of the columns)	278	1.00		H′ = 2.075	316	1.000		H′ = 1.753

[a] Abbreviations: n_i, density or number of the ith species; p_i, proportional abundance of the ith species.

Table 5. Calculation of Simpson's Dominance Index (D^{-1}) using weed species from the mineral soil meadow habitat.

	Mineral meadow[a]						
Weed species	n_i	$n_i - 1$	$n_i(n_i - 1)$	N	$N - 1$	$N(N - 1)$	$n_i(n_i - 1)/N(N - 1)$
Polygonum convolvulus L.	53	52	2756	278	277	77006	0.036
Amaranthus retroflexus L.	11	10	110	278	277	77006	0.001
Chenopodium album L.	15	14	210	278	277	77006	0.003
Brassica kaber (DC.) L. C. Wheeler	16	15	240	278	277	77006	0.003
Taraxacum officinale G. H. Weber ex Wiggers	36	35	1260	278	277	77006	0.016
Agropyron repens (L.) Beauv.	26	25	650	278	277	77006	0.008
Stellaria media (L.) Vill.	40	39	1560	278	277	77006	0.020
Medicago sativa L.	27	26	702	278	277	77006	0.009
Avena fatua L.	54	53	2862	278	277	77006	0.037
Σ (sum of the columns)	278						$D = 0.130$

[a] Abbreviations: n_i, density or number of the ith species; N, total number of individuals of all species.

The values of these diversity indices are relative, in the sense that there is no predetermined value of H' that determines whether or not a community is diverse; rather, these calculated values serve to compare diversities among communities. Thus, in the case above, the mineral meadow has greater species richness than the organic meadow. Richness can be visually estimated from a small set of data; a large sample requires the calculation of diversity indices.

For evenness, values of zero indicate that the community is extremely uneven (dominated by one species), whereas a value of one indicates that the community is absolutely even (maximum diversity exists, no one species dominates). The example above shows that the organic meadow has greater species evenness than the mineral meadow, even though the latter has greater species richness. The mineral meadow is dominated by one forb and two grass species: wild buckwheat (*Erigonum* spp.), quack grass [*Agropyron repens* (L.) Beauv.] and wild oat (*Avena fatua* L.); whereas the organic meadow is occupied evenly by forbs.

Simpson's Dominance Index (D). The formula (Equation 9) and calculations follow.

$$D = \Sigma\{[n_i(n_i - 1)]/[N(N - 1)]\} \qquad [9]$$

where: n_i = density or number of the ith species; and N = total number of individuals of all species in the community. By convention, Simpson's Dominance Index is usually written as the reciprocal value D^{-1}. The higher the index value, the more species evenness there is. A worked example is provided below (Table 5).

$$D_{(\text{mineral meadow})} = 0.130, \text{ therefore:}$$

$$D^{-1}_{(\text{mineral meadow})} = 1/0.130 = 7.69$$

Following similar calculations,

$$D_{(\text{organic meadow})} = 0.132, \text{ therefore:}$$

$$D^{-1}_{(\text{organic meadow})} = 1/0.132 = 7.58$$

In the example above, no one species dominates in either meadow, hence the values of D^{-1} are relatively high. They are corroborated by the Shannon–Weiner Diversity Index and Evenness calculations.

Between-Community Diversity. Between-community diversity, also known as "beta-diversity," is a measure of similarity or distinctiveness between communities within the landscape. It is meant to be used to compare communities from different areas or habitats within the defined landscape area. Similarity indices such the Sørensen and Steinhaus (Sørensen 1948) indices are used frequently to calculate between-community diversity. The Sørensen Coefficient Index is a function of the number of individuals of species common to communities that are evaluated: it is a similarity index. The Steinhaus Coefficient Index, on the other hand, is a function of abundance: it estimates the smallest abundance for each species established in different communities as a proportion of the average community abundance.

The Sørensen Coefficient Index (S_s). This index (S_s) is expressed as follows in Equation 10.

$$S_S = [2J/(a + b)] \times 100 \qquad [10]$$

Table 6. Indices used to calculate similarity between communities: the Sørensen Coefficient Index (S_s) and the Steinhaus Index (S_A). This hypothetical example compares the weed species found in two communities: the mineral soil meadow and the high organic soil meadow.[a]

	Mineral meadow abundance	Organic meadow abundance	Minimum abundance value
Weed species			
Polygonum convolvulus L.	53	0	0
Amaranthus retroflexus L.	11	41	11
Chenopodium album L.	15	58	15
Brassica kaber (DC.) L. C. Wheeler	16	48	16
Taraxacum officinale G. H. Weber ex Wiggers	36	59	36
Agropyron repens (L.) Beauv.	26	78	26
Stellaria media (L.) Vill.	40	0	0
Medicago sativa L.	27	32	27
Avena fatua L.	54	0	0
Total no. of individuals	278	316	131
Total no. of species	9	6	
No. of common species (J)	6		
Sum of the lower of the two abundances (W)	131		
Sørensen Coefficient Index (S_s):			
$S_s = [2J/(a + b)] \times 100$			
$S_s = [(2 \times 6)/(9 + 6)] \times 100$			
$S_s = [(12)/(15)] \times 100$			
$S_s = 0.8 \times 100$			
$S_s = 80\%$			
Steinhaus Index (S_A)			
$S_A = 2W/(A + B)$			
$S_A = (2 \times 131)/(278 + 316)$			
$S_A = 262/(594)$			
$S_A = 0.441$			

[a] Abbreviations: J, number of common species; W, sum of the lower of the two abundances; a, total number of individuals in community a; b, total number of individuals in community b; A, total number of individuals in population A; B, total number of individuals in population B.

where: J = the number of species common to each community; and $a + b$ = the sum of the total number of species in each community.

The Steinhaus Coefficient Index (S_A). This index (S_A) is shown in Equation 11:

$$S_A = W/[(A+B)/2] = 2W/(A+B) \qquad [11]$$

where: W = the sum of the lower of the two abundances of each species in the community; A = total number of individuals in population A; and B = total number of individuals in population B.

To calculate both indices, we will use the same set of data from Table 3 and assume that mineral and organic meadows are two distinct weed communities (Table 6). For both indices, we interpret their values on a scale from 0 (complete dissimilarity) to 1 (complete similarity).

The Sørensen Index suggests that the mineral and the organic meadow communities are quite (80%) similar (a value of 100% would indicate absolute similarity). A visual assessment, however, is possible in this case, and it shows that the two communities do differ; grasses and forbs are found in the mineral meadow, whereas the organic meadow is inhabited exclusively by forbs. This is why the Steinhaus Index is used. It accounts for differences in abundance, and hence is more accurate that the Sørensen Index. Other methods such as the Mountford's Index of Similarity (Mountford 1962; see also Wolda 1981) can resolve this issue as well. Note that one reason the Sørensen Index can be a bit misleading is its sensitivity to small samples, and in the above example, the sample size is small.

Methods to Evaluate Weed Associations and Composition: Multivariate Analysis

Statistical procedures in agricultural experiments generally entail highly controlled systems of exper-

iments. Here, a reduced number of factors, at limited ranges, are allowed to affect the subject (crops, weeds) being tested. Consequently, the range of responses, the variation of uncontrolled factors, and the experimental error are kept relatively small. In this context, statistical theory can be used to describe the pattern of response through a mathematical approximation of the distribution (e.g., a normal distribution).

In contrast, weed community data rarely fit model assumptions of normality even when they are obtained from planned experiments. Typical examples are weed surveys that entail hundreds of weed species across several sites and years. These types of studies assess responses of weed species to environmental and/or managerial factors. In these cases, weed species display a wider range of genetic diversity. As a consequence, the range of responses of weed species is far larger than the diversity found when dealing with major crop species. Moreover, the interactions that invariably occur among species or between weed species and environment or management variables can be very complex. For these types of data, the standard statistical procedures are clearly inappropriate. Consequently, given such a level of data complexity, it is of interest to try to structurally simplify the data set (i.e., reduce the dimensionality of the complexity), to identify similarities among species (classification), or among variables (grouping) in order to build and test hypotheses. Multivariate analysis is the statistical methodology that can elicit such information from complex data. It is the field of statistics that deals with the relationships among p variables measured on n objects or individuals. There are several multivariate techniques that can be used, depending upon the goal of the study and the type of data, such as principal component analysis, cluster analysis, canonical correlation, discriminant analysis, and factor analysis.

The basic observational unit for this type of data is usually a two-way table (matrix) of, for instance, weed species by sites or weed species by weed control methods. To avoid unnecessary complexity in describing the various numerical operations, the concise notation of matrix algebra is used. In this section, we only provide a brief review of concepts from matrix algebra that we use in our case studies. Summary details about matrix algebra are presented by Digby and Kempton (1987), and Khattree and Naik (2000), whereas further details with statistical viewpoints can be found in books by Harville (1997), Rao and Rao (1998), and Schott (2005).

Concepts from Matrix Algebra. *Observational Matrix*. Consider, for example, a case in which we have n weed species and on each of them are observed (same) p different characteristics (variables), or the species are measured at the same p different locations, say $x_1, x_2, \ldots x_p$. Then these data can be presented as an n by p matrix **X**:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots x_{1p} \\ x_{21} & x_{22} & \ldots x_{2p} \\ \vdots & & \\ x_{nl} & x_{n2} & \ldots x_{np} \end{bmatrix}$$

where X is the observation matrix, and the individual observations or measured variables are listed on each row. Each weed species can be viewed as a multivariate observation. The observations or measured variables are often correlated. It is a common practice to express each variable in such a way that it has zero mean and, optionally, unit variance: the process is called standardization. Thus, if $x_{i.}$ is the mean and s_i^2 the variance of the ith variable, x_{ij} can be standardized to:

$$y_{ij} = \left(x_{ij} - x_{i.}\right),$$

to set each variable to have zero mean, or as:

$$y_{ij} = \left(x_{ij} - x_{i.}\right)/s_i,$$

so that each variable also has unit variance,

where:

$$x_{i.} = 1/n \sum_{j=1}^{n} x_{ij} \quad (j \text{ is any of the } n \text{ weed species}) \quad [12]$$

Covariance Matrix. A large data set made of several variables, observed on several weed species (often hundreds), must be summarized in order to be understood. For univariate data, this is done through the use of basic descriptive statistics such as the mean and the variance. Likewise, for multivariate analysis, the population mean and the population covariance between variables can be determined. The description, however, is dramatically simplified by the use of matrix notations. Thus, for instance, suppose x_i was a random variable for which $x_{1,} \ldots x_p$ are possibly dependent. The covariance between any two variables x_i and x_j, denoted $\sigma_{ij} = \text{cov}\ (x_i, x_j)$ is calculated as:

$$\mathrm{cov}(x_i,x_j) = 1/n \sum_{k=1}^{n} (x_{ik} - x_{i.})(x_{jk} - x_{j.}) \quad [13]$$

and the population of variances (var) and covariances (cov) is displayed as a matrix **Σ:**

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1,x_2) & \ldots \mathrm{cov}(x_1,x_p) \\ \mathrm{cov}(x_2,x_1) & \mathrm{var}(x_2) \ldots & .\mathrm{cov}(x_2,x_p) \\ \vdots & \vdots & \vdots \\ \mathrm{cov}(x_p,x_1) & \mathrm{cov}(x_p,x_2) & \mathrm{var}(x_p) \end{bmatrix}$$

It should be noted that *cov* (x_i, x_i) = *var* (x_i), and the term c*ov* (x_i, x_j) is the (*i*, *j*)th entry in matrix **Σ.** The variance of the *i*th variable is at the *i*th diagonal place, whereas all the covariances are placed on the nondiagonal places. For this reason, matrix **Σ** is referred to as the variance–covariance matrix, or simply as the covariance matrix, and sometimes as the dispersion matrix.

The sum of the diagonal elements of **Σ** (read ''trace'' of **Σ),** written *tr*(**Σ**), is called the ''total variance'' and the determinant of **Σ** (|**Σ**|) is referred to as the ''generalized variance.'' Both are often taken as the overall measures of variability among the set of variables.

Correlation Matrix. It is advisable to start a multivariate analysis with a correlation matrix, instead of a covariance matrix, when the measurements recorded on the various variables are not on the same scale and variances are not of similar magnitude. If indeed the scale and variable variances are similar, it is suggested to use a covariance matrix (see Everitt 1989; and Jolliffe 2002, for discussion regarding the choice of correlation matrix over covariance matrix). An example of differing scales, for example, is the measurement of seed, root, and shoot weights of a weed species, which are likely to be in the order of milligrams, grams, and kilograms, respectively.

There is a relationship between the correlation (ρ_{ij}) and the covariance between two variables x_i and x_j as calculated with Equation 14:

$$\rho_{ij} = \mathrm{cov}(x_i,x_j) / \left[\mathrm{var}(x_i)\mathrm{var}(x_j)\right]^{1/2} \quad [14]$$

where ρ_{ij} is the Pearson's population correlation coefficient between x_i and x_j. The population correlation matrix is then defined as:

$$\rho = \left(\rho_{ij}\right) = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

Transpose, Diagonal, Vector, Orthogonal, and Identity Matrices. Let **A** be a (*m* by *n*) matrix; its ''transpose,'' noted **A′**, is the matrix **B** (*n* by *m*) whose elements b_{ij} are defined as: $b_{ij} = a_{ji}$. A ''diagonal matrix'' is a symmetric ($x_{ij} = x_{ji}$) matrix with $x_{ij} = 0$ for all $i \neq j$; its only nonzero elements occur on the leading diagonal. A ''vector'' is a matrix with only one row (row vector) or one column (column vector). Finally, a matrix **A** (*n* by *n*) is said to be an ''orthogonal'' matrix if:

$$\mathbf{A'A} = \mathbf{AA'} = \mathbf{I_n} \quad [15]$$

where $\mathbf{I}_n$ is the identity matrix, a diagonal matrix with all diagonal elements equal to 1.

Matrix Decomposition. Many of the multivariate techniques use matrix decomposition, which is the expression of a matrix as the product of two or more matrices. One of the two decomposition methods often used is the ''spectral decomposition'' of a symmetric matrix. Let **A** be a symmetric matrix of order *n* [**A** (*n* by *n*)]. It is demonstrated that **A** can be written as:

$$\mathbf{A} = \mathbf{P\Lambda P'}$$

Where **P** is an orthogonal matrix of order *n*, and **Λ** is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. The scalars (algebraic numbers) are called the ''eigenvalues'' (or ''latent roots'') of **A.** Each column of **P** is an ''eigenvector (or ''latent vector'') of **A**.

Principal Component Analysis. Principal component analysis (PCA) is one of the oldest and most widely used multivariate methods. It mainly serves as an exploratory investigative tool. The purpose of multivariate analysis is to lessen the complexity of large data set by reducing its dimensionality. This is often achieved by reducing the number of variables or creating new variables that are functions of the original variables. In the case of PCA, the new

uncorrelated variables are linear combinations of the original ones. These new functions are called "principal components," and the statistical procedure of finding them without sacrificing most of the information contained in a data set is called "principal component analysis."

Determination of the Principal Components. Let's suppose $\mathbf{\Sigma}$ is the covariance matrix of p variables x_1, … , x_p measured on n weed species. According to Equation 13, $\mathbf{\Sigma}$ can be "spectrally" decomposed as:

$$\mathbf{\Sigma} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

In full, this can be written as:

$$\begin{bmatrix} \mathbf{\Sigma} & & & \\ c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix} = \begin{bmatrix} \mathbf{P} & & & \\ l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{bmatrix}$$

$$\times \begin{bmatrix} \mathbf{\Lambda} & & & \\ \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} \begin{bmatrix} \mathbf{P}' & & & \\ l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & l_{p2} \\ \vdots & \vdots & & \vdots \\ l_{1p} & l_{pp} & \cdots & l_{pp} \end{bmatrix}'$$

The scalars λ_1; λ_2; … ; λ_n are the eigenvalues (or latent roots) of the covariance or correlation matrix; whereas each column of $\mathbf{P}$ is an eigenvector (or latent vector) of the covariance or correlation matrix. The elements of each eigenvector represent the coefficients of the corresponding principal component, which is a linear combination of the original variables. From the example above, the eigenvalue λ_1 has for corresponding eigenvector $\mathbf{p_1}' = (l_{11}, l_{21}, \ldots l_{p1})$ which drives the first principal component ξ_1:

$$\xi_1 = l_{11}x_1 + l_{21}x_2 + \ldots + l_{p1}x_p$$

The variance of the first principal component $\boldsymbol{\xi_1}$ is equal to λ_1 [var $(\boldsymbol{\xi_1}) = \lambda_1$] and its contribution to the total variance represented by the sum of the eigenvalues is given by the ratio $\lambda_1/\Sigma\ \lambda_i$.

The eigenvalue λ_k ($k \leq p$) has for corresponding eigenvector $\mathbf{p_k}' = (l_{1k}, l_{2k}, \ldots l_{pk})$ which drives the kth principal component ξ_k:

$$\xi_k = l_{1k}x_1 + l_{2k}x_2 + \ldots + l_{pk}x_p$$

Interpretations. The variance of the kth principal component $\boldsymbol{\xi_k}$ is equal to λ_k [var $(\boldsymbol{\xi_k}) = \lambda_k$] and its contribution to the total variance represented by the sum of the eigenvalues is given by the ratio $\lambda_k/\Sigma\ \lambda_i$.

The first principal component $\boldsymbol{\xi_1}$ has the highest variance among all linear combinations. The second principal component $\boldsymbol{\xi_2}$ has the second highest variance and is uncorrelated to the first. Similarly, the kth principal component $\boldsymbol{\xi_k}$ has the kth highest variance and is uncorrelated with all the other $(p - 1)$ principal components. Thus, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ and $var\ (\boldsymbol{\xi_1}) = \lambda_1 \ldots var\ (\boldsymbol{\xi_p}) = \lambda_p$. In summary, the eigenvalues of the covariance or correlation matrices represent the variances of the corresponding principal components. The sum of all eigenvalues is equal to the sum of all the diagonal elements of the covariance matrix ($tr\ \mathbf{\Sigma}$), which represents the total variance. Hence, the set of principal components cumulatively accounts for the total variance displayed by the data.

Like the eigenvalues, eigenvectors, namely their elements l_{ij}, which are the coefficients of the principal components, that is, the linear relationships between p variables, can have some interesting interpretations. Let x_i and y_j be the ith variable and the jth principal component, respectively. The correlation $corr\ (x_i, y_j)$ between x_i and y_j is:

$$corr(x_i, y_j) = l_{ji}\left[\lambda_j l\ \mathrm{var}(x_i)\right]^{1/2}$$

where l_{ji} is the coefficient of the ith variable of the the jth principal component, and λ_j represents the eigenvalues associated with the jth principal component. Equation 14 shows that the degree of association between a given variable in a principal component and that principal component is proportional to the coefficient associated with that variable. The higher the coefficient, the greater is the association between the variable and the principal component. In other words, the magnitude of a given coefficient is proportional to the contribution of the associated variable to the principal component.

Selection of the Number of Principal Components. The appropriate number of principal components to be selected and the choice of either covariance or

Table 7. SAS output showing correlation matrix.

Output Weed Survey

PCA Using Correlation Matrix

The PRINCOMP Procedure

Observations	*100*
Variables	*10*

Correlation Matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1									-
x_2	-	1								-
x_3	-		1							-
x_4	-			1						-
x_5	-				1					-
x_6	-					1				-
x_7	-						1			-
x_8	-							1		-
x_9	-								1	-
..										1

correlation matrix as the starting point for principal component analysis are the two issues that have been debated extensively in the literature without a real consensus. General guidance about the second issue has been provided above (Correlation Matrix). Regarding the selection of the number of principal components, three methods are commonly used: the ''scree diagram'' (graphical method); ''the size of the variance of the principal components'' (when the correlation matrix is the starting point); and the ''cumulative proportion of total variance.'' The latter, the most commonly used criterion, can be used irrespective of the type of matrix (covariance or correlation). The common practice in the literature is that a minimum percentage of total variation desired to be explained by the principal component analysis is predetermined, and the smallest number of principal components that satisfies this criterion is selected. In general, that minimum percentage is set at 90%.

Example Study. Suppose a weed scientist wishes to study the biology of the five most important weed species across a landscape that is the habitat of 10 weed species ($n = 10$) in total. A survey on weed abundance is carried out over 10 locations in the landscape ($p = 10$). A statistical method is required to measure the differences in weed density across the

Table 8. Eigenvalues of the correlation matrix.

	Eigenvalue	Difference	Proportion	Cumulative
1	5.334	3.112	0.5334	0.5334
2	2.222	1.099	0.2222	0.7556
3	1.823	0.399	0.1823	0.9379
4	0.523	1.300	0.0523	0.9902
5	0.0187	0.5043	0.00187	0.9920
6	0.0180	0.0007	0.00180	0.9938
7	0.0177	0.0003	0.00177	0.9955
8	0.0166	0.0011	0.00166	0.9972
9	0.0155	0.0011	0.00155	0.9988
10	0.0115	0.004	0.00115	0.99995

landscape. A table of weed mean densities will not be sufficient, because it will not report the interrelations among locations and it will not tell how weed density covaries among locations. In other words, the table of means cannot account for the distribution of weed species across the landscape. The principal component analysis procedure might help meet the weed scientist's goal.

Computer programs such as SPSS and SAS can be used to perform multivariate computations and tasks. For instance, in the case above, a SAS data set, designated as WEED, containing the raw data on the location variables x_1 to x_{10} can be created; and the PROC PRINCOMP statement can follow to achieve the PRINcipal COMponent analysis. With SAS, the default matrix is the correlation matrix; the covariance matrix is obtained by adding the option "COV" to the proc princomp statement. The detailed SAS program is as follows:

```
Title1 "Output Weed Survey";
data weed;
input x1 - x10 @@;
datalines;
x11 x12 x13 x14 x15 x16 x17 x18 x19 x110
x21  —  —  —  —  —  —  —  —  x210
 —   —  —  —  —  —  —  —  —   —
x101 x102 —  —  —  —  —  —  —  x1010
;
proc princomp data = weed;
var x1 — x10;
title2 'PCA using Correlation Matrix';
run;
OR:
proc princomp data = weed cov;
var x1 — x10;
title2 'PCA using Covariance Matrix';
run;
```

For the sake of illustration, let's assume SAS output displays the following correlation matrix, eigenvalues, and eigenvectors information (Tables 7, 8, and 9).

Interpretation

Case scenario 1. The goal of this study was to assess the significance of a weed infestation sampled across 10 locations within a defined landscape. These locations are surveyed for 10 weed species (variables x_i in this case). Species x_1, x_2, and x_6–x_{10} are dicotyledons; x_3–x_5 are monocotyledons; x_1, x_2, x_6, and x_7 are cooler-temperature-species, and x_3, x_4, x_5, x_8, x_9, and x_{10} are warmer-temperature species.

By examining the cumulative proportion of the variation explained by the principal components displayed in the output above, we see that three principal components are needed to account for at least 90% of the total variability. Further examination of the coefficients of the variables in these principal components show the first principal component ξ_1 with all its coefficients positive, that is, 0.458345, 0.481190, 0.392150, 0.366378, 0.345601, 0.401029, 0.465323, 0.312865, 0.432765, 0.289755 (see eigenvectors in Table 9). Thus, ξ_1 is given by:

$$\xi_1 = 0.46x_1 + 0.48x_2 + 0.39x_3 + 0.37x_4 + 0.35x_5 + 0.40x_6 + 0.47x_7 + 0.31x_8 + 0.43x_9 + 0.29x_{10} \quad [16]$$

The first principal component ξ_1, which accounts for 53% of the total variation, seems to measure the index of weed species significance across the 10 locations in the landscape. That is, ξ_1 would be an indicator of the level of infestation by the 10 weed species at each surveyed site of the landscape. All the coefficients are of similar magnitude, which suggests that abundances of the 10 weed species are in the same range across the landscape. The 10 surveyed sites can then be arranged in order according to the magnitude of the 10 scores derived from Equation 16.

Table 9. Eigenvectors.[a] (Values for Prin4–Prin10 were insignificant in this hypothetical case; i.e. close to zero or negative).

	Prin1	Prin2	Prin3	Prin4	—	Prin10
x_1	0.458345	0.056792	−0.597743	—	—	—
x_2	0.481190	0.114800	−0.221805	—	—	—
x_3	0.392150	−0.517139	0.069759	—	—	—
x_4	0.366378	−0.450765	0.565571	—	—	—
x_5	0.345601	−0.663735	0.551112	—	—	—
x_6	0.401029	0.285014	−0.001528	—	—	—
x_7	0.465323	0.543067	−0.145267	—	—	—
x_8	0.312865	0.123459	0.234857	—	—	—
x_9	0.432765	0.456432	0.289746	—	—	—
x_{10}	0.289755	0.143259	0.124789	—	—	—

[a] Abbreviations: Prin, principal component function; *x*, weed species.

The second principal component with the coefficients 0.056792, 0.114800, −0.517139, −0.450765, −0.663735; 0.285014, 0.543067, 0.123459, 0.456432, and 0.143259 seems to measure a difference between dicotyledons and monocotyledons. The coefficient for the variable x_1 (weed species 1) is as small as 0.056792; hence, this variable can be ignored in the definition of the second principal component. The third component seems to measure the difference between cool-temperature and warm-temperature weed species. Species 3 and 6 can be ignored from the definition of the third principal component.

Case scenario 2. The goal of this study was to assess the relative abundance of weed species in a given landscape in order to determine the three most important weed species in the landscape. In this case, the previous 2-way data matrix (*n* by *p*) would be reverted into a (*p* by *n*) data matrix, that is, the sites now become the variables and the weed species become the objects. Let's assume the 10 weed species' abundances are evaluated at 10 sites. Sites x_1, x_2, and x_6–x_{10} are located at a higher latitude; x_3–x_5 are located at a lower latitude; conventional tillage is the common practice in sites x_1, x_2, x_6, and x_7; and conservation tillage is the norm in sites x_3, x_4, x_5, x_8, x_9, and x_{10}. If we further assume, for the sake of illustration, that the results of the principal component analysis are the same as above, then the interpretation of the result would be slightly different.

The first principal component ξ_1, would measure the index of weed species abundance and distribution across the landscape. That is, ξ_1 would be an indicator of the importance of each weed species in the whole landscape. The second principal component would measure the difference between sites located at higher and lower altitudes, and the third principal component would measure the difference between sites in which conventional and conservation tillage are respectively practiced.

Canonical Discriminant Analysis. Canonical Discriminant Analysis (CDA) is a statistical technique similar to Principal Component Analysis and Canonical Correlation Analysis in its ability to reduce the dimensionality of a large data set. This technique is however specialized in ''discriminating'', that is, simultaneously studying the differences between two or more groups of objects with respect to several variables. Thus, CDA can be used to for interpreting the group differences or for classifying objects into groups. It can be particularly suited to weed community research. For instance, weed species can be used as variables to test whether a significant difference exists among predetermined groups such as tillage systems, weed control methods, ecosystems, etc. For the results to be statistically a good reflection of reality, certain assumptions need to be satisfied: (1) groups must be mutually exclusive; (2) a variable should not be a linear combination of other discriminating variables; (3) there should be similar group covariance matrices; and (4) each group is drawn from a population which has a multivariate normal distribution. This happens when each variable has a normal distribution about fixed values on all the others (Blalock 1979).

Deriving the Canonical Discriminant Functions. The nature of group differences can be studied through the use of canonical discriminant functions, which are linear combination of discriminating variables and have the following mathematical form:

$$f_{km} = u_o + u_1 X_{1km} + u_2 X_{2km} + \ldots + u_p X_{pkm}$$

where: f_{km} = the value (score) on the canonical discriminant function for object *m* in the group *k; k*

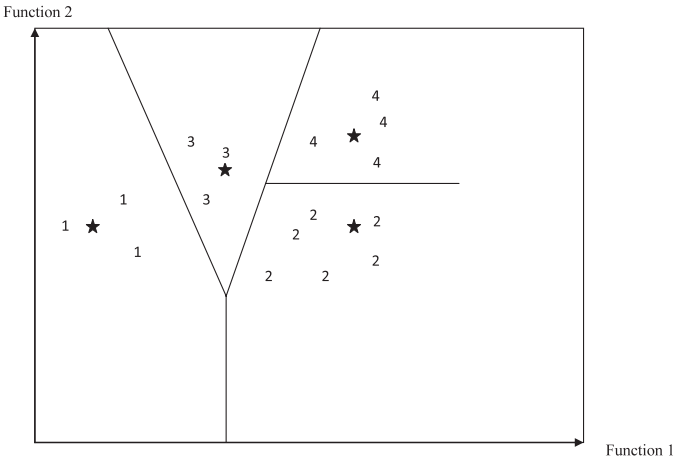


Figure 12. Hypothetical two-function plot of group centroids (stars) and individual objects (1, 2, 3, 4). The horizontal axis represents function 1, and the vertical dimension represents function 2.

$= \{1; 2; \ldots, g\}$; X_{ikm} = the value on discriminating variable X_i for object m in group k; and u_i = the ith coefficient on the canonical discriminant function. When there are p discriminating variables and g groups, the maximum number of unique canonical discriminant functions that can be analyzed is the smaller of the two numbers p and $(g - 1)$. These discriminant functions are automatically derived by various computerized statistical programs (e.g., SAS, SPSS).

Interpreting the Canonical Discriminant Functions. Once the canonical discriminant functions have been determined, their meaning can be interpreted by: (1) examining the relative positions of the objects and their group centroids (a group centroid being an imaginary point whose coordinates are the group's mean on each of the variable); and (2) studying the individual contribution of each discriminating variable to the discriminant function. When there is more than one discriminant function, the question of whether all of them are needed to describe the variability of the data set is examined as well.

Graphical separation of groups: the two-function plot. The location of group centroids and the observed objects can be plotted easily in a system of two axes (vertical and horizontal) represented respectively by two discriminant functions (Figure 12). Groups are distinct when their centroids are well-separated and there is no obvious overlap of the individual objects.

Standardized coefficients. Unstandardized coefficients in Equation 16 are obtained from original data. They indicate the absolute contribution of a variable in determining the discriminant score. This estimate of variable contribution, however, can be misleading in situations where the meaning of one unit change in the value of a variable is not the same from one variable to another, which is the case when the standard deviations of the various variables are not the same. To have a better idea about the relative contribution of each variable, it might be necessary to standardize the coefficient. This is achieved by subtracting the grand mean of the variable from each individual value and dividing the result by the standard deviation, so that the adjusted mean and standard deviation become zero and one, respectively. By examining the magnitude of the standard coefficient (ignoring the sign), the contribution of each variable to the score of the canonical discriminant function can be determined: the larger the magnitude, the greater is that variable's contribution.

Total structure coefficients. Although the standardized coefficients provide a way to estimate the variable's contribution to calculating the discriminant score, they can display a serious limitation in situations where two or more variables are highly correlated. In these cases, the variables share their contribution to the score. Their standardized coefficients might therefore be smaller than when only one of the correlated variables is used. Or, the contribution of one variable is partially cancelled by the opposite contribution of the other when the standardized coefficients of correlated variables are of opposite signs. One way to get around this limitation is to use total structure coefficients instead.

Structure coefficients are simple bivariate correlations between a single variable and a discriminant function. As such, they are not affected by the relationships with the other variables. A structure coefficient indicates how closely a variable is related to a canonical discriminant function. When the absolute value of the coefficient is close to or equal to one (+1 or −1), it means the function holds nearly the same information as the variable. When the structure coefficient is near zero, it means the variable and the function have very little in common.

The eigenvalues. In discriminant analysis, the canonical discriminant function with the largest eigenvalues (defined earlier) is the most powerful discriminator, that is, it is the function that will give

Table 10. Eigenvalues and measures of importance.

Canonical discriminant function	Eigenvalue	Relative percentage	Canonical correlation
1	10.65925	86.65	0.965
2	1.57915	12.83	0.874
3	0.06358	0.52	0.211

the best separation of the groups. Conversely, the function with the smallest eigenvalues is the weakest discriminator.

The relative percentage. The absolute numbers representing the eigenvalues cannot be interpreted directly. Their relative magnitudes when there is more than one function can help determine how much of the total discriminating power each function has. Such comparisons are obtained by converting the eigenvalues into relative percentages, that is, by dividing the sum of all eigenvalues (total discriminating power) into each individual eigenvalue. The proper interpretation of the relative percentage is that a function is strong or weak relative to others, that it is likely or unlikely to add further to the understanding of the differences between groups. However, the ''relative power'' of the first discriminant function does not necessarily translate into a strong association with the groups. For this reason, the canonical correlation is useful, because it tells how well a discriminant function is doing.

The canonical correlation coefficient. This coefficient measures the degree of association between the groups and the discriminant function. A value of zero denotes no relationship, whereas the maximally valued one indicates a perfect association. If the groups are distinct in the variables being analyzed, then all the correlation coefficients will be high. Both canonical correlation and relative percentage can be used to determine how many discriminant functions are meaningful and how useful they are in explaining group differences. The hypothetical statistics in Table 10 indicate that the first discriminant function contains 86.65% of the total discriminating power in this system of functions. Both the relative percentage and the correlation coefficient of the third function are very small, which suggests that function 3 is so unimportant that it lacks any research utility.

When data are from a population, then the number of functions and their importance are derived from the relative percentage and the canonical correlation. These two statistics completely characterize the degree of discrimination between the groups and the discriminating variables. When the data are from a sample, however, the question of statistical significance of the discriminant functions arises. The Wilks's lambda statistic can be used to settle this question.

Wilks's lambda statistic. The significance of discriminant functions is commonly tested indirectly through the examination of ''residual discrimination.'' A given function is not tested directly, but the residual discriminating power present in the system prior to the extraction of the function is examined. If, for instance, the residual discrimination is too small, then it is meaningless to extract more functions even when they mathematically exist. Wilks's lambda is a statistic (also known as the U statistic) that can be used to measure residual discrimination. It can also be converted to a test of significance of discriminant functions. There are several ways to calculate Wilks's lambda (Λ), one of which is:

$$\Lambda = \prod_{i=k+1}^{a} \frac{1}{1+\lambda_i}$$

Where k is the number of function already derived; λ_i the eigenvalue of the ith function; a the total number of eigenvalues; and the symbol Π indicates that the individual terms are to be multiplied in order to get the final product. Values of Wilks's lambda that are near zero indicate high discrimination. Values close to or equal to 1.0 indicate less or no discrimination.

The significance of Wilks's lambda can be tested by expressing it as a function of chi-square or F distributions (Klecka 1980). Thus, the chi-square formula can be written as follows:

$$\chi^2 = -\left[n. - \left(\frac{p+g}{2}\right) - 1\right] \log_e \Lambda_k$$

with $(p - k)$ $(g - k - 1)$ degrees of freedom; where: $n.$ = total number of individuals overall groups; g = number of groups; and p = number of variables.

Example study. For illustrative purposes, let's imagine that a weed scientist wants to explore

whether the choice of tillage systems affects the biology of weeds (morphology and physiology), which would suggest that tillage practices affect the competitiveness or aggressiveness of weed species. The objective of the study would be to develop a set of coefficients for certain morphological (shoot biomass, leaf area index) and physiological (growth rate, sucrose–phosphate synthase activity) traits, which could then be used to discriminate between three tillage systems: zero-, minimum-, and conventional-tillage. The study includes *m* weed species on which four discriminating variables, shoot biomass (SB), leaf area index (LAI), growth rate (GR), and sucrose–phosphate synthase activity (SPS) are to be measured.

If using the SAS program, the procedure to used would be PROC CANDISC, and the three tillage systems would be the CLASS variable TILLTYPE. The data file is named WEED. The following SAS program would perform the statistical analysis:

```
data weed;
Input tilltype $ sb lai gr sps;
if tilltype = 'zero tillage' then mark = 'Z';
if tilltype = 'minimum tillage' then mark = 'M';
if tilltype = 'conventional tillage' then mark = 'C';
proc candisc data = weed out = outcan bsscp pcov
   pcorr;
class tilltype;
var sb lai gr sps;
Title1 'Canonical Discriminant Analysis: Weed
   Science Society Data';
proc sort data;
by mark;
proc print data = outcan;
var tilltype mark can1 can2;
Title2 'Scores on Canonical Variables: Weed Science
   Society Data';
run;
symbol1 value = C;
symbol2 value = M;
symbol3 value = Z;
proc gplot data = outcan;
where mark = "C" or mark = "M" or mark = "Z";
plot can2*can1 = mark;
title1 h = 1.2 'Plot of the two Canonical
   Discriminant Function';
run;
```

Output and interpretation. The number of canonical discriminant functions is the smaller of the two numbers ($g - 1$) and p; where g = group number and p = number of variables. In this case, there are $(3 - 1) = 2$ canonical discriminant functions.

The options *bsscp, pcov, and pcorr* stand for "between sum of squares and cross-products," "pooled variance-covariance," and "pooled correlation," respectively. They result in respective matrices. In the absence of actual data, let's focus on hypothetical outputs of correlation coefficients, eigenvalues, and total structure coefficients (Tables 11 and 12).

Results from the pooled correlation matrix (PCORR) would have indicated the likelihood of correlation among the four discriminating variables: shoot biomass, leaf area index, growth rate, and SPS activity. Because growth rate is by definition the production of biomass per unit of time, shoot biomass is likely correlated to growth rate. Consequently, the total structure coefficient would be more appropriate as they are not affected by the relationships with the other variables (see paragraph "*Total structure coefficients*:, above).

The canonical correlation table displays the correlation coefficient between the two canonical discriminant functions (dependent variables) and the three tillage systems (independent variables). They are high for both functions (0.98 and 0.95), which would indicate a strong relationship between the three tillage systems and the two discriminant functions.

Next is the output of eigenvalues, which are the estimated variances of respective canonical discriminant functions. It can be observed that the first discriminant function can explain 76% of the total variance, and the remaining variance (24%) is explained by the second discriminant function. This does suggest that the two canonical discriminant functions should describe the data accurately, provided that their significance is confirmed by statistical tests under the assumption of multivariate normality. These tests are performed in the next table of the SAS output. The significance level of 0.0001 tells us that both canonical discriminant functions are highly significant.

From the table of structure coefficients, the two canonical discriminant functions f_1 and f_2 can be expressed as follows:

$$f_1 = 0.776548 * SB + 0.443273 * LAI + 0.558943 * GR + 0.0254398 * SPS$$

$$f_2 = 0.127549 * SB + 0.20054 * LAI + 0.501325 * GR - 0.009874 * SPS$$

Table 11. Hypothetical outputs of correlation coefficients, eigenvalues, and test of the Null Hypothesis.[a]

	Canonical correlation	Canonical correlations: Adjusted canonical correlation	Canonical correlations: Approximate standard error	Canonical correlations: Squared canonical correlation	Eigenvalues: Eigenvalues	Eigenvalues: Difference	Eigenvalues: Proportion	Eigenvalues: Cumulative	Test of HO: Likelihood ratio	Test of HO: Approximate F value	Test of HO: DF	Test of HO: Den DF	Test of HO: Pr > F
1	0.983567	—	—	0.967404	15.2954	105908	0.76477	0.76477	—	—	—	—	< 0.0001
2	0.952036	—	—	0.906372	4.7046		0.23523	1.0000	—	—	—	—	< 0.0001

[a] Abbreviations: HO, null hypothesis; df, degrees of freedom; Den df, denominator degree of freedom; Pr, probability.

Table 12. Total structure coefficients.[a]

Variable	Can1	Can2
sb	0.776548	0.127549
lai	0.443273	0.20054
gr	0.558943	0.501325
sps	0.0254398	−0.009874

[a] Abbreviations: Can, canonical discriminant function; sb, shoot biomass; lai, leaf area index; gr, growth rate; sps, sucrose–phosphate synthase activity.

The first canonical discriminant function f_1, dominated by the discriminating variable shoot biomass (SB), appears to represent growth and development of individual weed plant. The second canonical discriminant function f_2, largely influenced by the discriminating variable growth rate (GR), seems to represent a contrast between morphological traits and the physiological trait of SPS activity. In both functions, the coefficients for SPS are near zero, which means that the SPS activity and the discriminant functions have very little in common; as a result, the variable SPS activity could be dropped from both discriminant functions.

To see how effective the separation of the three tillage systems is, using the two canonical discriminant functions, a plot of the score on these two canonical discriminant functions was obtained using the SAS statements (Figure 13):

proc gplot data = outcan;
where mark = "C" or mark = "M" or mark = "Z";
*plot can2*can1 = mark;*

where the variable MARK takes values C, M, Z for the three tillage systems (conventional-,

Figure 13. Plot of the two canonical discriminant functions f_1 and f_2.

minimum-, and zero-tillage, respectively). The resulting plot presented above would suggest that the conventional-tillage system can be discriminated from the two conservation tillage practices (zero- and minimum-tillage). Additionally, the first canonical discriminant function would probably suffice to discriminate between conventional and the rest. Neither of the two functions or both, concomitantly, is able to separate between the two reduced-tillage methods.

Conclusions

Exploring how plant populations change over time in response to imposed selection pressures can be challenging. In this paper we have presented our views on field experiments and their analyses. We do not claim to have presented the one and only approach, but we hope that we have introduced you, the reader, to the critical topics that must be thought through carefully as you plan for a study in this area. By following the suggestions and information contained in this chapter, it is our hope that you will be successful in describing weed communities, species distributions and changes in plant diversity within the agroecosystem.

Literature Cited

Allen DW (2009) GIS Tutorial II. Redlands, CA: ESRI Press. 408 p

Blalock HM, Jr (1979) Social Statistics. New York: McGraw-Hill. 625 p

Booth BD, Murphy SD, Swanton CJ (2010) Invasive Plant Ecology in Natural and Agricultural Systems. 2nd edn. Cambridge, MA: CAB International. 299 p

Conroy MJ, Noon BR (1996) Mapping of species richness for conservation of biological diversity: conceptual and methodological issues. Ecol Appl 6:763–773

Cousins SH (1991) Species diversity measurement: choosing the right index. Trends Ecol Evol 6:190–192

Digby PGN, Kempton RA (1987) Appendix: Matrix algebra. Pages 193–203 *in* Multivariate Analysis of Ecological Communities. New York: Chapman and Hall

Everitt BS (1989) Statistical Methods in Medical Investigation. London: Edward Arnold. 195 p

Flanagan C, Jennings C, Flanagan N (1994) Automated GIS Capture. Pages 25–38 *in* Worboys MF, ed. Innovations in GIS 1. London, UK: Taylor and Francis

Forcella F, Harvey SJ (1988) Patterns of weed migration in northwestern U.S.A. Weed Sci 36:194–201

Gaston KJ (1991) How large is a species' geographic distribution? Oikos 61:434–437

Gaston KJ, Fuller RA (2009) The sizes of species' geographic ranges. J Appl Ecol 46:1–9

Gauch HG, Jr (1982) Multivariate Analysis in Community Ecology. Cambridge, UK: Cambridge University Press. 295 p

Goff FG, Dawson GA, Rochow JJ (1982) Site examination for threatened and endangered plant species. Environ Manag 6 (4):307–316

Gomez KA, Gomez AA (1984) Sampling in experimental plots. Chapter 15. Pages 532–561 *in* Statistical Procedures for Agricultural Research. 2nd edn. New York: John Wiley and Sons

Gorr WL, Kurland KS (2008) GIS Tutorial. 3rd edn. Redlands, CA: ESRI Press. 434 p

Harville DA (1997) Matrix Algebra from a Statistician's Perspective. New York: Springer-Verlag. 634 p

Huebner CD (2007) Detection and monitoring of invasive exotic plants: a comparison of four sampling methods. Northeast Nat 14 (2):183–206

Jolliffe IT (2002) Principal Component Analysis. New York: Springer-Verlag. 489 p

Kercher SM, Frieswyk CB, Zedler JB (2003) Effects of sampling teams and estimation methods on the assessment of plant cover. J Veg Sci 14:899–906

Khattree R, Naik D (2000) Concepts from matrix algebra, Pages 7–20 *in* Multivariate Data Reduction and Discrimination. Cary, NC: SAS Institute, Inc

Klecka RW (1980) Discriminant Analysis. Beverly Hills, CA: Sage Publications. 71 p

Lo CP, Yeung AKW (2007) Concepts and Techniques of Geographic Information Systems. 2nd edn. Cranbury, NJ: Pearson Education. 532 p

Magnussen S, Boyle TJB (1995) Estimating sampling size for inference about the Shannon–Weiner and the Simpson indices of species diversity. For Ecol Manag 78:71–84

Magurran AE (1988) Ecological Diversity and its Measurements. London, UK: Croom Helm. 179 p

Mayo T (1994) Computer-assisted tools for cartographic data capture. Pages 39–52 *in* Worboys MF, ed. Innovations in GIS 1. London, UK: Taylor and Francis

McClave TJ, Dietrich FH, II (1988) Estimation and test of hypothesis: single sample. Pages 317–402 *in* Statistics. 4th edn. San Francisco, CA: Dellen Publishing

McCune B, Grace JB (2002) Analysis of Ecological Communities. Gleneden Beach, OR: MjM Software Design. 304 p

Mitchell A (1999) The ESRI Guide to GIS Analysis. Volume 1: Geographic Patterns and Relationships. Redlands, CA: ESRI Press. 186 p

Mitchell A (2009) The ESRI Guide to GIS Analysis. Volume 2: Spatial Measurements and Statistics. Redlands, CA: ESRI Press. 238 p

Mountford MD (1962) An index of similarity and its application to classificatory problems. Pages 43–50 *in* Murphy PW, ed. Progress in Soil Zoology. London, UK: Butterworths

Pearman PB, Randin CF, Broennimann O, Vittoz P, van der Knaap WO, Engler R, Le Lay G, Zimmermann NE, Guisan A (2008) Prediction of plant species distribution across six millennia. Ecol Lett 11:357–369

Penskar MR (1991) Survey of the Ottawa National Forest for Endangered, Threatened, and Special Concern Species in Land Type Association 7 and Associated Ecological Land Type Phases. Lansing, MI: Michigan Natural Features Inventory, 21 p plus appendices

Rao CR, Rao MB (1998) Matrix Algebra and Its Applications to Statistics and Economics. Singapore: World Scientific. 535 p

Schlesinger RC, Funk DT, Roth PL, Myers CC (1994) Assessing changes in biological diversity over time. Nat Areas J 14:235–240

Schott JR (2005) Matrix Analysis for Statistics. New York: John Wiley and Sons. 480 p
Snedecor GW, Cochran WG (1967) Statistical Methods. 6th edn. Ames, IA: Iowa State University Press. 593 p
Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. K Dan Vidensk Selsk Biol Skr 5:1–34
Stiling P D (1999) Ecology: Theories and Applications. 3rd edn. Upper Saddle River, NJ: Prentice Hall. 638 p
Thomson DQ, Stuckey RL, Thompson EB (1987) Spread, Impact, and Control of Purple Loosestrife (*Lythrum salicaria*) in North American Wetlands. Fish and Wild Life Research Report No. 2. Washington, DC: U.S. Department of Interior, Fish and Wildlife Service. 55 p
Wilson JB, Steel JB, King WM, Gitay H (1999) The effect of spatial scale on evenness. J Veg Sci 10:463–468
Wolda H (1981) Similarity indices, sample size and diversity. Oecologia 50:296–302
Yorks TE, Dabydeen S (1998) Modification of the Whittaker sampling technique to assess plant diversity in forested areas. Nat Areas J 18:185–189

Received May 13, 2013, and approved October 31, 2013.