

TEACHING MULTIVARIATE STATISTICS TO ECOLOGISTS AND THE DESIGN OF ECOLOGICAL EXPERIMENTS TO STATISTICIANS: LESSONS FROM BOTH SIDES ®

Marti J. Anderson
University of Auckland
New Zealand

Multivariate data in ecological applications most often occur in the form of counts of species abundances in assemblages, where each species is a variable. These data do not generally conform to traditional statistical assumptions, and so special approaches and methods are needed in this context. Statisticians need to be informed about these special problems with ecological data. In addition, the rationale for complex experimental designs that is a trademark of most ecological studies needs to be well understood by applied statisticians in this area. On the other hand, successful approaches for teaching ecologists about the use of multivariate statistics include sticking to the conceptual, rather than the mathematical. I provide here an overview of the methods that have helped teaching across these two disciplines, including a general approach for the use of novel non-parametric methods in the analysis of ecological community data.

INTRODUCTION

Multivariate analysis in ecological applications most often involves the investigation of interactions or effects in natural systems on the fauna and/or flora of a particular area. Such ecological assemblages consist of many interacting species, and each species is generally treated as a separate response variable in the analysis. Thus, a typical multivariate data set in ecology usually has counts of several species (columns, say) at each of several sites or in each of several observation units (rows) (e.g. Orłóci, 1975; Gauch, 1982; Digby & Kempton, 1987; Jongman, ter Braak & van Tongeren, 1995, Legendre & Legendre, 1998).

There are three main categories of multivariate analysis that are common in ecology: (i) clustering, (ii) ordination and (iii) statistical tests of hypotheses. Clustering is generally concerned with the identification of clusters or groups of observations (or species) and is most often used for the generation of hypotheses in the beginning of a study or for exploratory purposes. Although this is a vast area of research, I shall not be concerned with it here, but rather shall focus on the situation where the ecologist has a particular hypothesis in mind and has gone out to sample natural systems in a structured way or has done a manipulative experiment of some kind. In this context, the main agenda is to produce an ordination in order to visualize patterns in the multivariate cloud of data (by reducing dimensions in some way) and also to rigorously test explicit hypotheses concerning those patterns by reference to *a priori* groups or relationships with predictors, such as environmental variables.

TEACHING MULTIVARIATE STATISTICS TO ECOLOGISTS

There are several important points to be kept in mind when teaching multivariate statistical methods to ecologists. First of all, most ecologists may be expected to have a basic background in univariate statistics (e.g. basic sampling of populations, *t*-tests, ANOVA, *P*-values and the like), usually from a general undergraduate course in biometry. However, very few, if any, practicing ecologists will have seen much in the way of multivariate analysis in any undergraduate program in biology or ecology, in most places in the world (although, hopefully, this is slowly changing). Ecologists generally become interested in multivariate analysis because they *already have multivariate data* (with some accompanying, although often vague, hypotheses) and they need to use multivariate methods to obtain and understand their results. In addition, with many ecologists, there is an intrinsic fear and loathing of mathematics. This of course is not always the case, but it bears remembering that ecology, as a science, grew largely out of natural history and descriptive observation (e.g. Thoreau, 1908). A more quantitative scientific framework of field measurement, sampling, experimental protocol and statistical analysis developed much later (e.g. Connell, 1961; Green, 1979; Underwood, 1981). Experience with teaching multivariate statistics to biologists and ecologists (as a graduate-level course at the

University of Sydney and to professional ecologists in external short-courses in many places) suggests the following rules of thumb:

1. Keep the mathematics to a minimum. If you say the word “eigenvalue” most ecologists will head for the hills! Matrix algebra is a lurking nightmare from high school for most practicing ecologists. This doesn’t mean that the algebra is beyond them, or that mathematics is not important, it just means that a typical mathematical treatment of the topic, like that found in many good multivariate statistical texts, such as Mardia, Kent and Bibby (1979), should be provided as extra rather than required reading, and supporting texts that can help should be suggested, such as Searle (1982).

2. Use conceptual notions that ecologists can relate to, such as specific kinds of differences in community structure, and relate these to geometrical representations in ordinations. Different methods and the use of different measures of dissimilarity or distance, for instance, can be described and visualized conceptually, even without the mathematical algorithm being articulated. Lucid examples of this conceptual approach are found in Clarke and Green (1988) and Clarke (1993), who present non-parametric methods for analysis of community data in an extremely intuitive and accessible manner. This means the ecologist will have the ability to use and interpret multivariate analyses without getting bogged down or lost in the mathematics.

3. In describing ordination methods, draw examples in two dimensions. Two dimensions can be drawn and seen easily, but nevertheless do represent a multivariate (bivariate) system. Thus, many complex methods can be understood conceptually or geometrically simply by explicitly drawing on paper (or on a chalkboard) how the method would reduce the system from two dimensions down to one. This is a particularly good way to see, also, that ordination methods do not actually change the data cloud in multivariate space, they just change the way (or the angle) that we choose to look at that cloud.

4. Try to draw analogies with univariate methods that ecologists would already be familiar with. For example, show how Hotelling’s T^2 is a direct multivariate extension of Student’s t -test. Students usually respond well when they can connect something new to something familiar that they have seen before and used with confidence.

5. Emphasize that multivariate analysis is not hypothesis-free. Some researchers collect large loads of data (i.e. they “go out and measure everything”) in a natural field-based situation in the vain hope that multivariate analyses will somehow magically provide the tool they need both to ask the right questions and then to answer them. This is clearly not the case. The logic of ecological experiments that should be applied concerning univariate hypotheses and tests (e.g. Underwood, 1990) applies equally to multivariate hypotheses. Field experiments, whether involving some structured sampling of natural patterns or involving manipulative treatments, need to be planned with care and logical thought (e.g. Hurlbert, 1984; Andrew & Mapstone, 1987).

6. Ecologists should be challenged (or even forced!) to articulate the particular hypothesis (or hypotheses) of interest for the study. Why did they collect the data in the first place? Articulation of a reasonable ecological multivariate hypothesis is (at least) half of the battle.

7. Perhaps the most important thing to engender in students is an understanding of the link between the logic of the underlying hypothesis (including the experimental design and how data were collected) and the nature of the appropriate multivariate analysis to be done. There will be many choices to make (e.g. choice of transformation, distance measure, ordination method, and testing strategy), but only some subset of multivariate methods will actually be appropriate for a given set of hypotheses and with data of a particular kind. Within such a subset, different methods will tend to give similar results. Getting students to recognize which multivariate analyses would be reasonable and appropriate in different situations is not as difficult as it sounds and can be achieved at a conceptual level, without students knowing much mathematics.

TEACHING THE DESIGN OF ECOLOGICAL EXPERIMENTS TO STATISTICIANS

Now let’s put the shoe on the other foot. It might appear at first glance that my purpose in the previous section is to denigrate ecologists for their lack of statistical knowledge. Not so! The variability and complexity of ecological systems demands for much more than a casual understanding of statistics and probability, and most ecologists demonstrate this understanding successfully on a regular basis in their work. Indeed, experience in teaching graduate students of

statistics some of the important things that make the analysis of multivariate ecological data unique has demonstrated that problems in communication or understanding can definitely come from either camp. There are several key features to ecological systems that most statisticians are completely unaware of and which may cause serious problems for correct inference unless taken into account. In a graduate-level course at the University of Auckland, I have attempted to introduce some of these special considerations to students of statistics, such as:

1. Multivariate ecological data, generally consisting of counts of species abundances, are usually extremely non-normal. Often, the distribution of counts of any individual species will include many zeros and will demonstrate right (positive) skewness, with varying levels of aggregation (Taylor, 1961; McArdle, Gaston & Lawton, 1990; Gaston & McArdle, 1994). In addition, unfortunately, traditional multivariate methods, unlike their univariate counterparts, are generally not robust to departures from normality (Mardia, 1971; Olson, 1974; Johnson & Field, 1993). This means that the multivariate methods commonly taught as part of undergraduate and graduate courses in multivariate statistics largely serve as little more than a curiosity to ecologists. This may come as something of a shock and disappointment to many statisticians, who may enjoy the elegance of the likelihood-based inferential methods (e.g. Seber, 1984; Rencher, 1998).

2. Virtually all ecological data are obtained in the field, from natural systems. These systems are naturally stochastic, with many interacting variables. There is no such thing as a “laboratory” here, where extraneous variation may be “controlled.” It is important to relay the message to statisticians that natural temporal and spatial variability are *always* present in ecological studies. The consequence of this is that ecological sampling programs or experimental designs need to take this into account. This is usually done by either: (a) replicating the experiment in time and space at an appropriate scale (e.g. Hurlbert, 1984) or (b) explicitly modeling the spatial or temporal structure in the data (e.g. Legendre & Fortin, 1989; Dutilleul, 1993). Incorrect inferences and conclusions can be drawn from ecological experiments that fail to take into account natural temporal and spatial variability.

3. Statisticians rarely consider that our perceptions of patterns in ecological systems are strongly scale dependent. Arbitrary decisions concerning the size and shape of the sampling unit, the scale, frequency, number and extent of sampling can all have major consequences on statistical inference (e.g. Andrew & Mapstone, 1987; Wiens, 1989; Thrush et al., 1997).

4. As a consequence of the natural temporal and spatial variability in ecological systems, the majority of the experimental designs used by ecologists are going to include random factors, often in nested hierarchies (e.g. Underwood & Chapman, 1998) or mixed models (e.g. Beck, 1997). Many of these designs are an anathema to students of statistics, who do not generally encounter much in the way of mixed models, at least in their undergraduate careers. Most undergraduate statistical courses in experimental design are concerned with the analysis of fractional factorial designs and deliberate confounding (ignoring interaction terms), as are used in engineering and agricultural applications. Such designs are basically never used in ecology. Why? Because spatio-temporal interactions with treatment effects are often the terms of greatest interest in ecological studies! For example, generality of effects may be inferred from the lack of significant interactions of fixed main effects (such as predation) with random spatial or temporal factors (Beck, 1997). Also, in the assessment of environmental impact, it is the statistical significance of the *interaction* between the contrast of impacted versus control locations and the contrast of samples taken before versus after the impact that indicates whether a significant impact has occurred (e.g. Green, 1979; Underwood, 1993).

5. Statisticians need to be made aware of the special problems in the analysis of multivariate ecological data. The combination of misbehaving data, complex experimental designs, ever-present spatial and temporal variability and the fact that patterns are scale-dependent can cause mayhem for rigorous statistical inferences if they are ignored. It is all too easy for statisticians to consider that problems and data sets in ecology are like any other system and so apply traditional methods (ignoring violations of assumptions) or suggest traditional designs (ignoring interaction terms or random effects), resulting in erroneous conclusions.

A NON-PARAMETRIC APPROACH FOR MULTIVARIATE ANALYSIS IN ECOLOGY

As a consequence of the known problems with using traditional multivariate statistical methods with ecological data, we have developed novel non-parametric methods (McArdle & Anderson, 2001; Anderson, 2001a; Anderson & McArdle in review; Anderson & Robinson in review; Anderson & Willis in review). These methods are flexible in that they allow the analysis to be based on any distance or dissimilarity measure of choice, but they also allow partitioning according to any linear model of interest for the experimental design, including tests of interaction terms. *P*-values are then obtained by use of an appropriate permutational strategy for each particular term in the model (e.g. Manly, 1997; Anderson, 2001b).

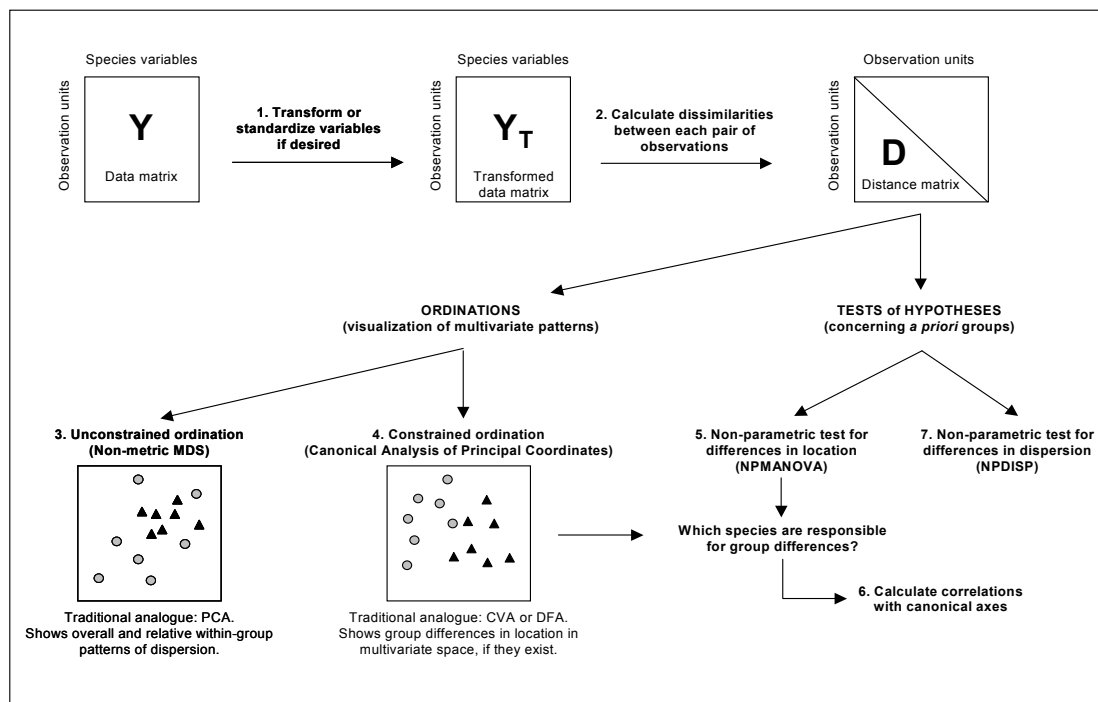


Figure 1. Schematic diagram of a general approach for multivariate analysis of ecological data, using novel non-parametric methods of ordination and hypothesis testing.

In the spirit of Clarke (1993) and Clarke and Green (1988), I suggest here a general strategy for the analysis of ecological multivariate species abundance data by reference to *a priori* hypotheses in a given experimental design. I shall restrict attention to the situation where the hypothesis concerns differences among groups, such as in an ANOVA design. However, these methods and tests can be done just as easily for the situation where one posits relationships between multivariate data and continuous predictor variables (such as environmental variables). The steps are shown schematically in Figure 1 and include:

1. Choose a transformation and/or standardization (if desired) to apply to the data, given the hypothesis and the nature and relative scales of the species variables.

2. Choose and apply a distance or dissimilarity measure appropriate for the analysis (e.g. see Legendre & Legendre, 1998 for a useful summary).

3. Do an unconstrained ordination on the basis of the distance matrix only, using non-metric or metric multi-dimensional scaling (Gower, 1966; Kruskal & Wish, 1978). This will give a visual representation of the overall pattern of dispersion and the relative within-group dispersions. The traditional analogue (in Euclidean space) is principal component analysis (PCA).

4. Do a constrained ordination by reference to the hypothesis, using canonical analysis of principal coordinates (also called generalized discriminant analysis based on distances) (Anderson & Robinson in review; Anderson & Willis in review). This method will find the axis that best separates groups (or that has maximal correlation with some continuous predictor variables) in the space defined by the particular distance measure chosen. The traditional analogue (in Euclidean space) is discriminant function analysis (DFA) or canonical variate analysis (CVA).

5. Test for differences in location among groups, using NPMANOVA and an appropriate permutation method (e.g. Manly, 1997; Anderson, 2001a, 2001b).

6. Investigate the species (original variables) that may be responsible for group differences (if they occurred) by investigating correlations with canonical axes (Anderson & Willis in review).

7. Test for differences in relative within-group dispersions (Anderson & McArdle in review). Note that this test is going to be strongly affected (not surprisingly) by the choice of transformation and also by the choice of dissimilarity measure at the outset. It is also important to note that such a test may be important in its own right, not just as a companion to tests for differences in location. For example, changes in relative multivariate dispersion have been suggested as indicators of environmental stress (Warwick & Clarke, 1993; Chapman, Underwood & Skilleter, 1995).

DISCUSSION

To maintain the nexus between ecology and statistics, as for any enduring relationship, if it is to last, communication is the key. Although this paper has focused on the specific issues that make multivariate analysis in ecology a challenge (from the perspective of either the statistician or the ecologist), I imagine that similar problems occur in other fields using multivariate analysis, such as psychology, economics, or medicine. The non-parametric methods developed for use in ecology and outlined above have flexibility and generality that will allow them to be used readily in these other fields as well.

ACKNOWLEDGEMENTS

I thank John Harraway for inviting me to write this paper. Computer programs (including user notes and examples) for performing the novel multivariate methods given in Figure 1 can be downloaded from <http://www.stat.auckland.ac.nz/PEOPLE/marti>.

REFERENCES

- Anderson, M.J. (2001a). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26, 32-46.
- Anderson, M.J. (2001b). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58, 626-639.
- Anderson, M.J., & McArdle, B.H. (in review). A new non-parametric test for homogeneity of multivariate dispersions. *Ecology*.
- Anderson, M.J., & Robinson, J. (in review). Generalised discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics*.
- Anderson, M.J., & Willis, T.J. (in review). Canonical analysis of principal coordinates: a new ecologically meaningful approach for constrained ordination. *Ecology*.
- Andrew, N.L., & Mapstone, B.D. (1987). Sampling and the description of spatial pattern in marine ecology. *Oceanography and Marine Biology Annual Review*, 25, 39-90.
- Beck, M.W. (1997). Inference and generality in ecology: current problems and an experimental solution. *Oikos*, 78, 265-273.
- Chapman, M.G., Underwood, A.J., & Skilleter, G.A. (1995). Variability at different spatial scales between a subtidal assemblage exposed to the discharge of sewage and two control locations. *Journal of Experimental Marine Biology and Ecology*, 189, 103-122.
- Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18, 117-143.
- Clarke, K.R., & Green, R.H. (1988). Statistical design and analysis of a biological effects study. *Marine Ecology Progress Series*, 46, 213-226.
- Connell, J.H. (1961). The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. *Ecology*, 42, 710-723.
- Digby, P.G.N., & Kempton, R.A. (1987). *Multivariate analysis of ecological communities*. London: Chapman and Hall.
- Dutilleul, P. (1993). Spatial heterogeneity and the design of ecological field experiments. *Ecology*, 74, 1646-1658.

- Gaston, K.J., & McArdle, B.H. (1994). The temporal variability in animal abundances: measures, methods and patterns. *Philosophical Transactions of the Royal Society of London B*, 345, 335-358.
- Gauch, H.G. (1982). *Multivariate analysis in community ecology*. Cambridge: Cambridge University Press.
- Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325-338.
- Green, R.H. (1979). *Sampling design and statistical methods for environmental biologists*. New York: Wiley.
- Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54, 187-211.
- Johnson, C.R., & Field, C.A. (1993). Using fixed-effects model multivariate analysis of variance in marine biology and ecology. *Oceanography and Marine Biology Annual Review* 31, 177-221.
- Jongman, R.H.G., ter Braak, C.J.F., & van Tongeren, O.F.R. (Eds.) (1995). *Data analysis in community and landscape ecology* (2nd edn). Cambridge: Cambridge University Press.
- Kruskal, J.B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: Sage Publications.
- Legendre, P., & Fortin, M.-J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, 80, 107-138.
- Legendre, P., & Legendre, L. (1998). *Numerical ecology*, 2nd English edition. Amsterdam: Elsevier B.V.
- Manly, B.F.J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*, 2nd edition. London: Chapman and Hall.
- Mardia, K.V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 58, 105-121.
- Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate analysis*. London: Academic Press.
- McArdle, B.H., & Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82, 290-297.
- McArdle, B.H., Gaston, K.J., & Lawton, J.H. (1990). Variation in the size of animal populations: patterns, problems and artifacts. *Journal of Animal Ecology*, 59, 439-454.
- Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894-908.
- Orlóci, L. (1975). *Multivariate analysis in vegetation research*. The Hague: W. Junk B.V.
- Rencher, A.C. (1998). *Multivariate statistical inference and applications*. New York: Wiley.
- Searle, S.R. (1982). *Matrix algebra useful for statistics*. New York: Wiley.
- Seber, G.A.F. (1984). *Multivariate observations*. New York: Wiley.
- Taylor, L.R. (1961). Aggregation, variance and the mean. *Nature*, 189, 732-735.
- Thoreau, H.D. (1908). *Walden, or, Life in the woods*. London: Dent.
- Thrush, S.F., Pridmore, R.D., Bell, R.G., Cummings, V.J., Dayton, P.K., Ford, R., Grant, J., Green, M.O., Hewitt, J.E., Hines, A.H., Hume, T.M., Lawrie, S.M., Legendre, P., McArdle, B.H., Morrisey, D., Schneider, D.C., Turner, S.J., Walters, R.A., Whitlatch, R.B., & Wilkinson, M.R. (1997). The sandflat habitat – scaling from experiments to conclusions. *Journal of Experimental Marine Biology and Ecology*, 216, 1-9.
- Underwood, A.J. (1981). Techniques of analysis of variance in marine biology and ecology. *Oceanography and Marine Biology Annual Review*, 19, 513-605.
- Underwood, A.J. (1990). Experiments in ecology and management: their logics, functions and interpretations. *Australian Journal of Ecology*, 15, 365-389.
- Underwood, A.J. (1993). The mechanics of spatially-replicated sampling programs to detect environmental impacts in a variable world. *Australian Journal of Ecology*, 18, 99-116.
- Underwood, A.J., & Chapman, M.G. (1998). A method for analyzing spatial scales of variation in composition of assemblages. *Oecologia*, 107, 570-578.
- Warwick, R.M., & Clarke, K.R. (1993). Increased variability as a symptom of stress in marine communities. *Journal of Experimental Marine Biology and Ecology*, 172, 215-226.
- Wiens, J.A. (1989). Spatial scaling in ecology. *Functional Ecology*, 3, 385-397.