Annu. Rev. Ecol. Syst. 1990. 21:129–66

MULTIVARIATE ANALYSIS IN ECOLOGY AND SYSTEMATICS: PANACEA OR PANDORA'S BOX?

Frances C. James

Department of Biological Science, Florida State University, Tallahassee, Florida 32306

Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, New York 14853

KEY WORDS: multivariate analysis, data analysis, statistical methods

INTRODUCTION

Multivariate analysis provides statistical methods for study of the joint relationships of variables in data that contain intercorrelations. Because several variables can be considered simultaneously, interpretations can be made that are not possible with univariate statistics. Applications are now common in medicine (117), agriculture (218), geology (50), the social sciences (7, 178, 193), and other disciplines. The opportunity for succinct summaries of large data sets, especially in the exploratory stages of an investigation, has contributed to an increasing interest in multivariate methods.

The first applications of multivariate analysis in ecology and systematics were in plant ecology (54, 222) and numerical taxonomy (187) more than 30 years ago. In our survey of the literature, we found 20 major summaries of recent applications. Between 1978 and 1988, books, proceedings of symposia, and reviews treated applications in ecology (73, 126, 155, 156), ordination and classification (13, 53, 67, 78, 81, 83, 90, 113, 121, 122, 159), wildlife biology (33, 213), systematics (148), and morphometrics (45, 164,

0066-4162/90/1120-0129$02.00

Table 1 Applications of multivariate analysis in seven journals, 1983–1988. In descending order of the number of applications, the journals are *Ecology*, 128; *Oecologia*, 80; *Journal of Wildlife Management*, 76; *Evolution*, 72; *Systematic Zoology*, 55; *Oikos*, 41; *Journal of Ecology*, 35; and *Taxon*, 27.

Principal components analysis	119
Linear discriminant function analysis	100
Cluster analysis	86
Multiple regression	75
Multivariate analysis of variance	32
Correspondence analysis	32
Principal coordinates analysis	15
Factor analysis	15
Canonical correlation	13
Loglinear models	12
Nonmetric multidimensional scaling	8
Multiple logistic regression	7
	514

200). For the six-year period from 1983 to 1988 (Table 1), we found 514 applications in seven journals.

Clearly, it is no longer possible to gain a full understanding of ecology and systematics without some knowledge of multivariate analysis. Or, contrariwise, misunderstanding of the methods can inhibit advancement of the science (96).

Because we found misapplications and misinterpretations in our survey of recent journals, we decided to organize this review in a way that would emphasize the objectives and limitations of each of the 12 methods in common use (Table 2; Table 3 at end of chapter). Several books are available that give full explanations of the methods for biologists (53, 128, 148, 159, 164). In Table 3, we give specific references for each method. In the text we give examples of appropriate applications, and we emphasize those that led to interpretations that would not have been possible with univariate methods.

The methods can be useful at various stages of scientific inquiry (Figure 1). Rather than classifying multivariate methods as descriptive or confirmatory, we prefer to consider them all descriptive. Given appropriate sampling, 6 of the 12 methods can also be confirmatory (see inference in Table 2). Digby & Kempton (53) give numerous examples of applications that summarize the results of field experiments. Most often the methods are used in an exploratory sense, early in an investigation, when questions are still imprecise. This exploratory stage can be a very creative part of scientific work (206, pp. 23–24). It can suggest causes, which can then be formulated into research hypotheses and causal models. According to Hanson (86), by the time the

Table 2 General objectives and limitations of multivariate analysis

Objectives	Codes to Procedures (see Table 3)
1. Description	All
2. Prediction	MR. LDFA, MLR
3. Inference	MR, MANOVA, LDFA, FA, MLR, LOGL
4. Allocation	LDFA
5. Classification	LDFA, MLR, CLUS
6. Ordination	LDFA, PCA, PCO, FA, CANCOR, COA, NMDS

Limitations:

1. The procedures are correlative only; they can suggest causes but derived factors (linear combinations of variables) and clusters do not necessarily reflect biological factors or clusters in nature.
2. Because patterns may have arisen by chance, their stability should be checked with multiple samples, null models, bootstrap, or jackknife.
3. Interpretation is restricted by assumptions.
4. Automatic stepwise procedures are not reliable for finding the relative importance of variables and should probably not be used at all.

theoretical hypothesis test has been defined, much of the original thinking is over. In the general scientific procedure, descriptive work, including descriptive applications of multivariate analysis, should not be relegated to a status secondary to that of experiments (28). Instead it should be refined so that research can proceed as a combination of description, modelling, and experimentation at various scales (106).

The opportunities for the misuse of multivariate methods are great. One reason we use the analogy of Pandora's box is that judgments about the results based on their interpretability can be dangerously close to circular reasoning (124, pp. 134–136; 179). The greatest danger of all is of leaping directly from the exploratory stage, or even from statistical tests based on descriptive models, to conclusions about causes, when no form of experimental design figured in the analysis. This problem is partly attributable to semantic differences between statistical and biological terminology. Statistical usage of terms like "effect" or "explanatory variable" is not meant to imply causation, so the use of terms like "effects" and "roles" in titles of papers that report descriptive research (with or without statistical inference) is misleading. Partial correlations and multiple regressions are often claimed to have sorted out alternative processes, even though such conclusions are not justified. "If . . . we choose a group of . . . phenomena with no antecedent knowledge of the causation . . . among them, then the calculation of correlation coefficients, total or partial, will not advance us a step toward evaluating the importance of the causes at work" (R. A. Fisher 1946, as quoted in reference 54, p. 432).

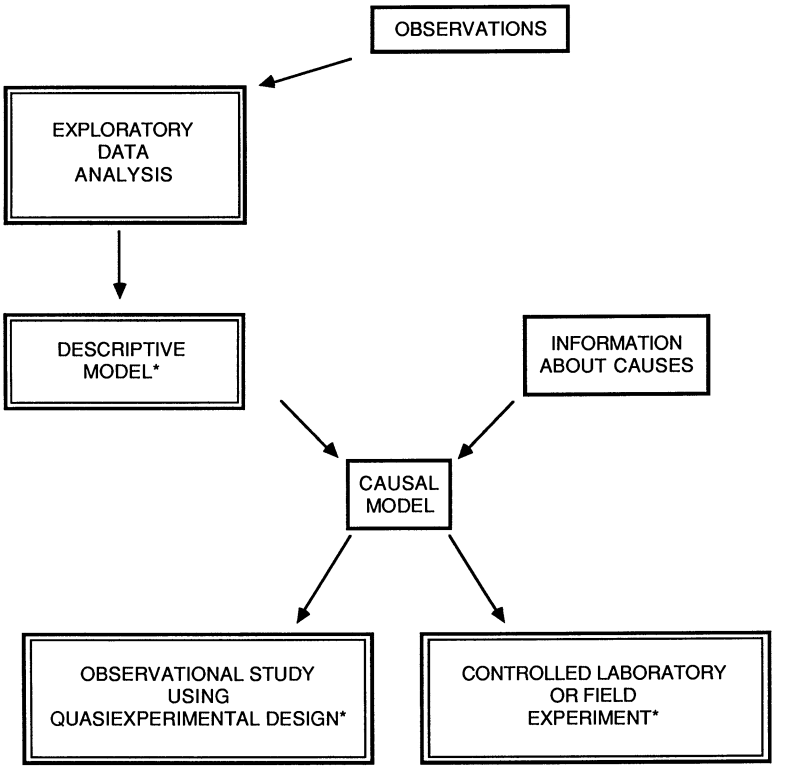

Figure 1 General research procedure showing stages (double boxes) at which exploratory and inferential* (confirmatory) multivariate analysis may be appropriate (modified from 106).

Although this idea is familiar to biologists, it seems to get lost when they enter the realm of multivariate work.

The objective of the present review is to help the researcher navigate between the Scylla of oversimplification, such as describing complex patterns with univariate analyses (147), and the Charybdis of assuming that patterns in data necessarily reflect factors in nature, that they have a common cause, or, worse, that statistical methods alone have sorted out multiple causes.

Present understanding of the role of multivariate analysis in research affects not only the way problems are analyzed but also how they are perceived. We discuss three particularly controversial topics, and we realize that not all researchers will agree with our positions. The first is the often-cited "problem" of multicollinearity, the idea that, if correlations among variables could be removed, one could sort out their relative importance with multivariate analysis. The problem here is a confusion between the objectives of the

method and the objectives of the researcher. Second, in the sections on analysis and ordination in plant ecology, we discuss the special problems that arise with indirect ordinations, such as the cases where the data are the occurrences of species in stands of vegetation. The arch pattern frequently seen in bivariate plots is not an artifact of the analysis; it is to be expected. Third, in the section on morphometrics, we explain why we argue that shape variables, which we define as ratios and proportions, should be studied directly. Of course the special properties of such variables require attention. We do not treat cladistics or the various software packages that perform multivariate analyses. In the last section, we give examples of how some basic concepts in ecology, wildlife management, and morphometrics are affected by the ways in which multivariate methods are being applied.

SUMMARY OF METHODS: OBJECTIVES LIMITATIONS, EXAMPLES

Overview

It is helpful to think of multivariate problems as studies of populations of objects about which information for more than one attribute is available (48, 169). One can describe the pattern of relationships among the objects (individuals, sampling units, quadrats, taxa) by ordination (reduction of a matrix of distances or similarities among the attributes or among the objects to one or a few dimensions) or by cluster analysis (classification of the objects into hierarchical categories on the basis of a matrix of inter-object similarities). In the former case, the objects are usually displayed in a graphic space in which the axes are gradients of combinations of the attributes. Principal components analysis is an ordination procedure of this type. It uses eigenstructure analysis of a correlation matrix or a variance-covariance matrix among the attributes. Principal coordinates analysis is a more general procedure in the sense that it starts with any type of distance matrix for distances among objects. Both principal components analysis and principal coordinates analysis are types of multidimensional scaling. Nonmetric multidimensional scaling uses the ranks of distances among objects, rather than the distances themselves. Correspondence analysis is an ordination procedure that is most appropriate for data consisting of counts (contingency tables). In this case, the distinction between objects and attributes is less relevant because they are ordinated simultaneously. Factor analysis is similar to principal components analysis in that it uses eigenstructure analysis, usually of a correlation matrix among attributes. It emphasizes the analysis of relationships among the attributes. Canonical correlation reduces the dimensions of two sets of attributes about the same set of objects so that their joint relationships can be studied.

When the objects fall into two or more groups, defined a priori, the