MULTIVARIATE TECHNIQUES IN LIMNOLOGY

by

A. L. Sheldon

Department of Zoology
University of Montana
Missoula, Montana

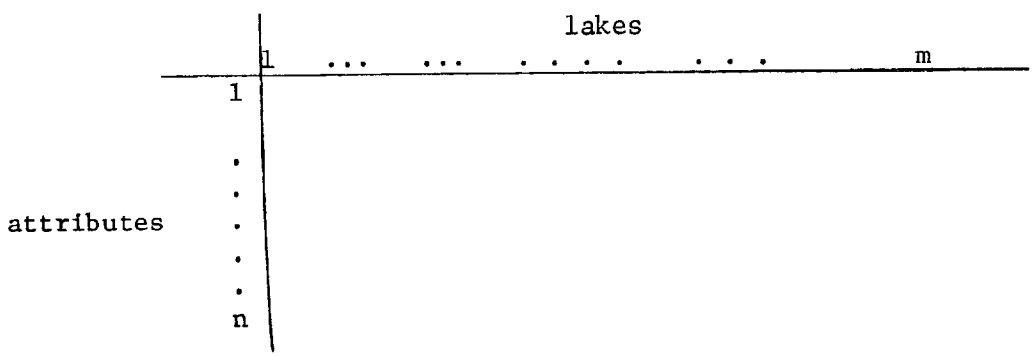
MULTIVARIATE TECHNIQUES IN LIMNOLOGY

INTRODUCTION

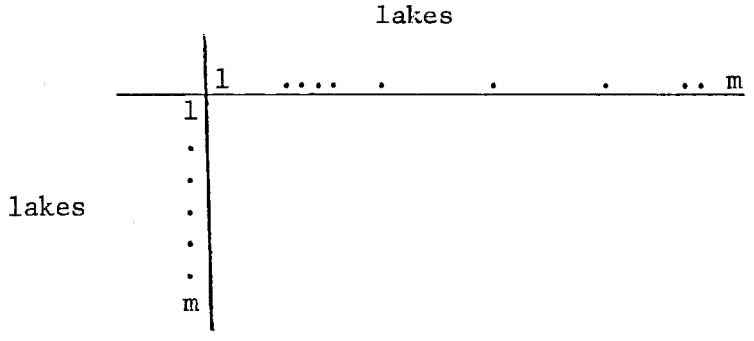
The classification problem in limnology is no different in principle from the problems of biological systematics or library science, and the remarks of Gilmour and Walters (1964) are pertinent to any of these fields. Multivariate techniques do little to resolve the philosophical problems inherent in classificatory work. Indeed, if all the thinking is done by the computer and the output accepted as revealed truth, multivariate techniques are no more than a complex but rapid method of attaining a dubious synthesis of questionable interpretation. This statement is not intended as a total condemnation of multivariate statistics in limnology or any other field. The limnologist or engineer can make good use of multivariate methods if he understands the limitations and idiosyncrasies of the various techniques.

THE BASIS OF MULTIVARIATE CLASSIFICATIONS

Any water body can be described by an infinite set of attributes -- morphometric, chemical, esthetic, economic and biological. Ignoring, for the moment, our criteria for the relevance and importance of characteristics, we may arrange the information in a lake x attribute matrix:



in which the attributes may be binary, ranked or continuous variables. Given such a matrix we may wish to group either the attributes or the lakes. Principal components reduction, canonical analysis and factor analysis (Shannon, this report) are appropriate techniques for the study of attribute groups and will not be discussed here. The first step in classifying or grouping the lakes is the construction of an m x m matrix of similarities or differences.



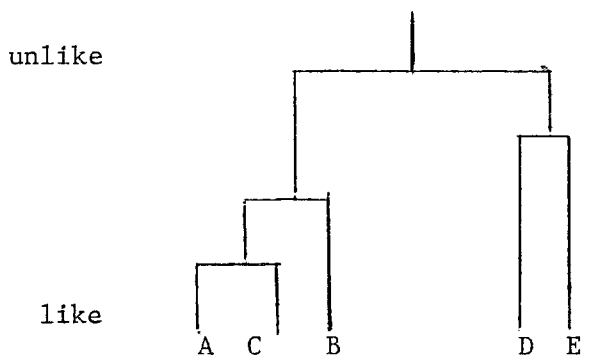
A multitude of similarity and difference measures have been proposed (Sokal and Sneath, 1963). Correlation and Euclidean distance are commonly used and, for binary data, various contingency measures. This is not the place for a critical discussion of all possible similarity indices, but it should be noted that "shape" coefficients (correlation) and "size" measures (Euclidean distance) lead to different results. For binary data, measures which consider negative matches as evidence of similarity (correlation) differ from those which do not (per cent similarity). Thus we are faced with a broad choice of alternatives at an early state of the classification process. Numerical classifications are not unique and the user must make a choice from an array of similarity methods and grouping strategies.

METHODS

Shannon (this report) has discussed a number of useful multivariate techniques. Two additional methods which have been used in ecology and other fields are hierarchic methods and the ordination procedure of Gower (1966).

Lance and Williams (1967) have reviewed a number of hierarchic methods. Many variations are possible but the following exemplifies one general method.

Given individuals (lakes) A, B, C, D and E characterized by multiple attributes, compute a similarity matrix. Scan the matrix and make a decision of the type: A and C are more alike than any other pair. Form the group AC, compute its average characteristics over all attributes and compute the similarity of AC with all individuals or groups remaining. Repeat the process until all individuals have been incorporated. The end product can be displayed as a dendrogram.



The major advantages of hierarchic methods are their simplicity and computational ease. Drawbacks include the tendency to generate residual groups and the distortion inherent in reducing a multidimensional situation to a two-dimensional tree. Like a decorative "mobile", the entire tree is free to pivot at each intersection, so the dendrogram

does not indicate that A or C is closer to B. Individual fusions do not use the information in the similarity matrix very efficiently. Certain strategies may generate "reversals" in which groups fuse at a lower level than did their constituent individuals. In spite of these drawbacks, hierarchic schemes are useful for rapid sorting of date and may be the only feasible method for large arrays such as a series of plankton samples (Brown, 1969).

The ordination method developed by Gower (1966) has much in common with principal components reductions and other standard techniques. The basic similarity matrix may be of the correlation or distance type. Relationships are projected on a series of orthogonal axes and the original inter-individual distances are preserved. Ordination retains the original information much better than do the hierarchic methods. However, storage requirements and running time are greatly increased. In some cases the added information may not warrant the effort involved. (Webb et al, 1967) have compared methods applicable to certain types of ecological data. A related paper (Williams et al, 1969) is of special interest since it deals with multivariate analysis of successional states analogous to eutrophication.

FLIES IN THE OINTMENT

It is obvious that there is no single multivariate method for classifying lakes, watersheds or plankton samples. Various combinations of measures and grouping techniques may be appropriate for different kinds of data and particular problems. Selection of a technique is largely a matter of informed judgment and experience with a variety of methods.

The real problem in multivariate classification is not the method but the original data. All such techniques are designed to minimize within-group variances or distances and maximize those between groups. This dependence on the variance requires that the attributes be given in some standardized units. Transformations such as the logarithmic will alter the variances and the classification. The most serious problem is redundancy of characters. Suites of redundant attributes can dominate the analysis. The problem lies in distinguishing between attributes which have been measured several times under different names e.g. lake "size" as area, mean depth, maximum depth, and those correlated by some important causal mechanism e.g. phosphorus, chlorophyll. Principal components reduction of the attributes identifies such correlated characters but does not tell one what to do with them. Factor analysis (rotation) increases the problem. In neither case are the eigenvalues a valid measure of importance. Presence-absence data for plant and animal species are free of the redundancy problem since there is a logical dividing line between the attributes and different lakes with many species in common can be considered similar. Even here problems may arise. If the species are members of groups of different size e.g. diatoms and copepods, the resultant classification will be dominated by diatoms and tell little about copepods unless the two groups are highly concordant.

Hierarchic techniques will usually isolate rare but unusual individuals in the residual group. In ordination procedures, relative abundance of individuals in confounded with the degree of difference and a rare but peculiar type may not appear in the first few vectors. Thus a single eutrophic or transitional lake in a sample of oligotrophic lakes might

not be detected.

The classical multivariate techniques, including the discriminant function, assume that individuals not included in the original analysis, are part of the same statistical universe. In limnology we very rarely sample, or even define, the universe under consideration. For this reason, I feel it is much safer to regard the sample as a closed system. Multivariate methods can be used to portray relationships within the sample but inferences about the rest of the world should be cautious ones.

USES OF MULTIVARIATE TECHNIQUES

Multivariate techniques are useful in processing and simplifying large quantities of information. The "arbitrary" nature of the sorting rules may generate groups and relationships which the more flexible human mind might never detect.

Multivariate classifications can be used in experimental design and analysis. There would seem to be some virtue in analyzing causal mechanisms in systems where some variables are constant in fact rather than adjusted by regression techniques.

A "good" classification is one which allows predictions to be made about variables not used in its construction. Unique or valuable lakes will be studied and managed on an individual basis but other cases such as the 15,000 lakes of Wisconsin will require some extrapolation of diagnostic and management techniques. Perhaps this is the area where miltivariate techniques will have their greatest impact.

REFERENCES

Brown, S. D. (1969). Grouping Plankton Samples by Numerical Analysis. Hydrobiologia, 33, 289-301.

Gilmour, J. S. L., and S. M. Walter. (1964). Philosophy and Classification. Vistas in Botany (Ed. W. B. Turrill) 4, 1-22.

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

Lance, G. N., and W. T. Williams. (1967). A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems. Computer Journal, 9, 373-380.

Sokal, R. R., and P. H. A. Sneath. (1963). Principles of Numerical Taxonomy. W. H. Freeman: San Francisco.

Webb, L. J., J. C. Tracey, et al, (1967). Studies in the Numerical Analysis of Complex Rain-Forest Communities. I. A Comparison of Methods Applicable to Site/Species Data. J. Ecol., 55, 171-191.

Williams, W. T., et al. (1969). Studies in the Numerical Analysis of Complex Rain-Forest Communities. III. The analysis of Successional Data. J. Ecol., 57, 515-535.