MERLEWOOD RESEARCH AND DEVELOPMENT PAPER

ISSN 0308-3675

Number 74



# NUMERICAL CLASSIFICATION AND CLUSTER ANALYSIS IN ECOLOGY: A REVIEW

P. J. A. Howard

R & D 77/74
December 1977

# 1. INTRODUCTION

Classification involves the recognition of similarities between, and the grouping of, objects and organisms. As a mental activity, this is probably considerably older than the writings of the ancient Greeks, the source of the science of classification. Indeed, in the form of pattern recognition, it may be fundamental to the way in which human beings and living organisms perceive the world around them.

In all sciences, as data accumulate, the necessity for simplification becomes apparent. A classification may have more than one purpose, but the paramount purpose is to describe the relationships of objects to each other, and to simplify the relationships so that general statements can be made about classes of objects. An important distinction is between monothetic and polythetic classifications. Monothetic classifications are those in which the classes established differ by at least one property which is uniform among the members of each class. In polythetic classifications, taxa are groups of individuals or objects that share a large proportion of their properties, but do not necessarily agree in any one property. A corollary of polythetic classification is the requirement that many properties be used to classify objects. However, once a classification has been established, few characters are generally necessary to allocate objects to the proper taxa. Classifications based on many properties will be general, they are unlikely to be optimal for any single purpose, but might be useful for a great variety of purposes. By contrast, a classification based on few properties might be optimal with respect to those properties, but would be unlikely to be of general use (Sokal, 1974).

Hence, classification of a data set results in a reduction of the amount of information that is necessary to describe the data, but, if the classification is efficient, there is little or no reduction in the amount of information contained in the data. Furthermore, classifications that describe relationships among objects in nature should generate hypotheses, possibly the main scientific justification for the exercise.

Much classificatory work in various branches of science has aimed to describe what is known as the 'natural system'. This is a difficult and controversial concept involving a variety of philosophical considerations which will not be pursued here (see e.g. Jardine and Sibson, 1971; Sneath and Sokal, 1973; Sokal, 1974).

Attempts to find satisfactory breaks in continuous data have led to similar principles and procedures being developed independently in various fields, and a body of general classification theory and methodology has been rapidly developing. Sokal (1974) stated that, in classification, theory has frequently followed methodology, and has been an attempt to formalize and justify the classificatory activity. In other instances, classificatory systems have been set up on a priori logical or philosophical grounds, and the methodology has been tailored subsequently to fit the principles. Both approaches have their advantages and drawbacks; modern work tends to reflect an interactive phase in which first one and then the other approach is used, but neither principles nor methodology necessarily dominates.

It is easy to perceive structure in data when the structure and discontinuities are obvious, but such a situation is not typical. Much of what we observe in nature changes continuously in one property or another, but not necessarily with equally steep gradients for each property. It is such cases which give the taxonomist (or ecological classifier) the greatest problems in deciding where or how to draw boundaries, or even whether boundaries should be drawn at all.

The development of numerical methods in taxonomy has had several effects. What used to be an intuitive art has been formalized into a quantitative science. The increasing availability of digital computers means that it has become practicable to explore the use of a wide range of numerical techniques, and this has attracted the attention of statisticians and mathematicians, with the consequent development of a wide variety of methods and their application to a variety of problems in different fields. This has not been an unmixed blessing, as there now exists a bewildering variety of numerical techniques, the properties of many being not fully known. Numerical methods can help the taxonomist investigate the structure of his data, but the results of the analyses still need to be interpreted. It is perhaps worth bearing in mind the three questions: Why do it? How do you do it? When you have done it, what does it mean? (Dr. A. J. Wilmott, pers. comm.).

The aim of this paper is to clarify some of the issues involved in the use of numerical methods in classification, particularly with regard to the search for structure in ecological data. A theoretical approach is taken because it is necessary to understand the theory of the methods in order to understand how they should be used and what their limitations are. Indeed, the same is true for multivariate methods in general. It is assumed that the reader is familiar with some multivariate theory, at least for the more widely used ordination techniques. Ecological classification is much less well-developed than is the taxonomy of organisms. This is due partly to the diversity of interests of ecologists, and partly to the nature of ecological data, which do not lend themselves to easy classification. Furthermore, the appropriate methods in taxonomy of organisms are not necessarily the most appropriate for ecological use (Clifford and Stephenson, 1975).

## 2. THE NATURE OF THE DATA

### 2.1 Types of data

Data consist of attribute scores, and there are many different kinds of attributes. These are discussed by, e.g. Clifford and Stephenson (1975). The most common kinds of attributes are: (i) Binary, e.g. presence-absence; (ii) Disordered multistate, e.g. colour; (iii) Ordered multistate, e.g. rare, common, abundant; (iv) Meristic, e.g. number of petals; (v) Continuous, i.e. measures on a continuous scale.

These attribute categories are not distinct, they depend to some extent on the sampling procedure, and data in one form can be converted to another. Differences of opinion exist on the value of binary data in ecological work, but the consensus of opinion seems to be that other data are preferable (Clifford and Stephenson, 1975, p. 39). In most branches of ecology, the tendency is to regard dominance (by some measure) as important. Results of analyses using data with numerical values are more informative than those using binary data. For example, Williams et al (1973) found that while plant species presence-absence was adequate in a simple study involving only eight sites, for ten sites there was "some advantage" in using numbers, while for 80 sites, quantitative data were distinctly preferable. Barkham (1968) found quantitative data to be more informative than presence-absence data in a study of the vegetation of Cotswold beechwoods.

The use of binary data in ecology can only be justified if it is difficult to obtain anything else, or if there is a declared lack of interest in the information which is lost by using binary data instead of, say, continuous data. Similarly, the conversion of meristic or continuous data to binary data is usually unsatisfactory. Thinking of this in terms of a normal distribution being arbitrarily divided into two sections, division along the mean leads to all entities on either side having identical binary scores. However, there may be instances when continuous data have properties which make conversion to another form logical. One example is if a variable can take a wide range of values, but has a concentration at one value (usually zero), as in counts of parasites on a host. It is usually better to regard such a variable as discrete, and to score it as if it were a few groups, such as zero, low, medium, high. It should be noted that if the aim of the analysis is to find a useful or meaningful grouping of the data, a coarsely-grouped variable may exert a disproportionate influence on the result (Marriott, 1974).

### 2.2 Selection of attributes

In ecological studies, there may or may not be a priori grounds for selecting attributes, and some attributes may be selected intuitively. There is likely to be some limitation on the types of attribute which can be used, due to practical difficulties in their measurement.

However the attributes are chosen, it is necessary to recognize that certain kinds of attributes are regarded as inadmissible in a numerical study (Jardine and Sibson, 1971; Sneath and Sokal, 1973). The different kinds of correlated attributes are particularly important in this respect. It is thus necessary to make some sort of initial check on the data.

Marriott (1974) discussed the problems of binary variables in cluster analysis. He noted that the selection of variables, important in any multivariate procedure, is paramount in the case of binary variables. If there is any inherent structure in the data, it should reveal itself in the dependencies between the variates. As a corollary, any variates that are non-independent for reasons not connected with an underlying grouping should be excluded from the analysis. For binary variables, the problem of deciding if there is more than one group, and if so how many and how they should be divided, is not easy. The justification for multimodality as a criterion is less clear in the case of binary variates than in that of continuous variates. In general, multimodality depends in a complicated way on the probabilities associated with each dichotomy in a $2^p$ contingency table for p binary variates.

## 2.3 Weighting of the attributes

The question of whether, or how, to weight data is an important problem in taxonomy, and specialists in different groups of organisms will have their own ideas about the importance of different attributes. One solution to this problem is to state the basis of weighting, so that the reader may judge the merits of the case (Clifford and Stephenson, 1975). Sneath and Sokal (1973) considered that equal weighting is desirable; it can be defended on several independent grounds, and is probably the only practical solution.

Jardine and Sibson (1971) concluded that certain kinds of weighting which taxonomists use intuitively are, in fact, incorporated in the calculation of K-dissimilarity, while certain of the other kinds of weighting and correlation which taxonomists have discussed were shown to be relevant to the selection of attributes, rather than to the calculation of dissimilarity and analysis of dissimilarity coefficients once attributes have been selected.

## 2.4 Possible structure in the data

As the underlying theme in numerical classification is the search for discontinuities in the data, it is important to think about what types of structure may be present in the data.

The first possibility is that there is only one group and all the individuals belong to it (e.g. Fig. 1). We may or may not know the nature of the distribution, but we cannot assume that it is normal. The data may contain more than one group. If there are real discontinuities, the groups will be separate and distinct. This situation presents no problems, the problems occur if the tails of the frequency distributions overlap (e.g. Fig. 2). Examples with a greater degree of overlap can be visualized, with the centres of the distributions moving closer together. In the extreme, this leads to a complex distribution which may appear to be unimodal.

This type of structure is visualized by plant ecologists. Sørensen (1948) suggested that the various types of vegetation are often so insensibly merged as to form a sliding scale, but that in a limited area under investigation it can be considered to be homogeneous with as much approximation to that mathematical concept as can be found in nature. Whittaker (1970, 1972) developed the continuum theory and showed that one might expect each species to find a different niche on an environmental gradient (e.g. Fig. 3). If the distribution patterns of species are completely continuous, it becomes impossible to delineate communities or

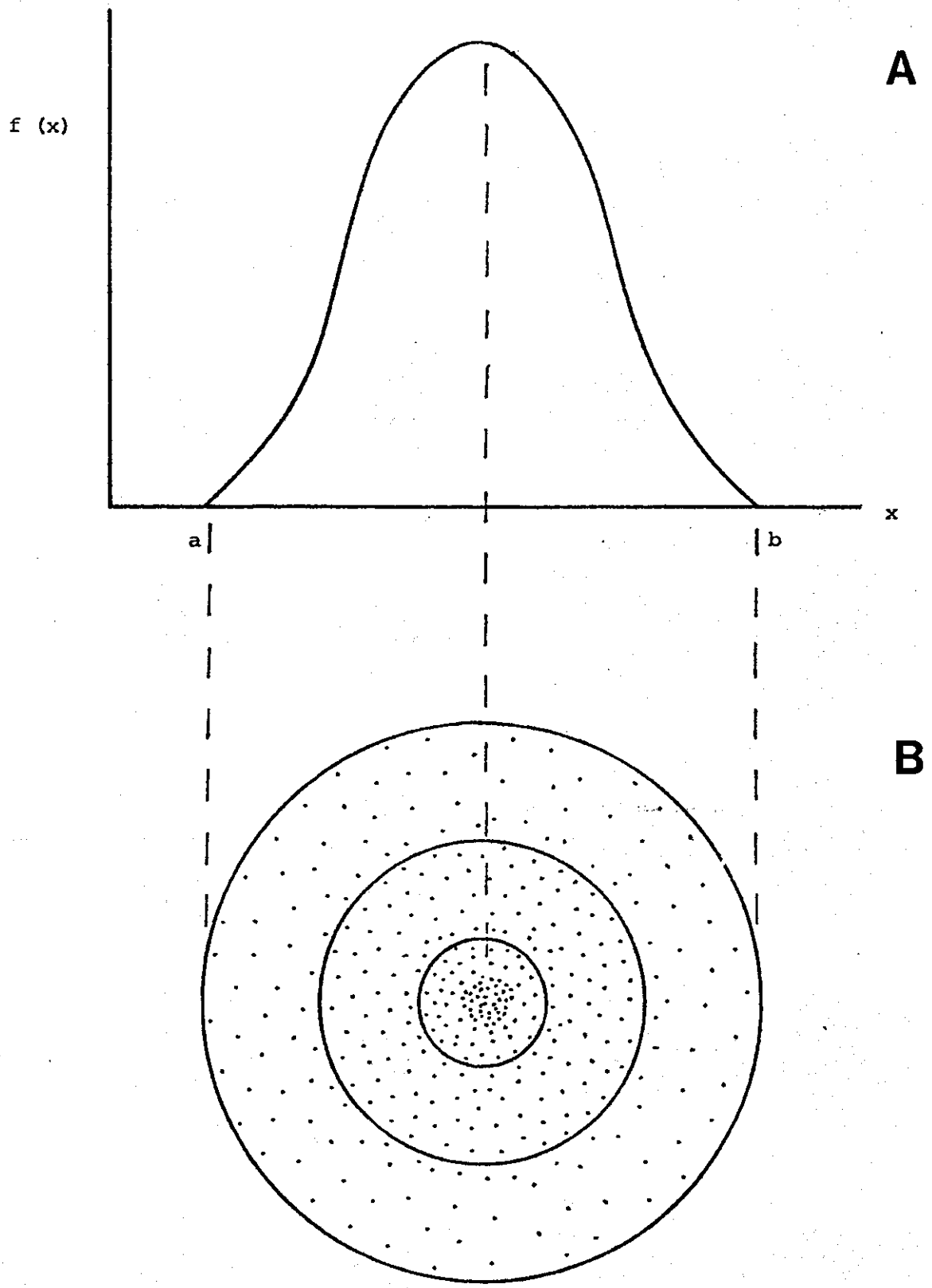


Figure 1. There is only one group, and all the OTU's belong to it. A is the graph of the continuous frequency (probability density) function f (x) of the data. This function is zero outside some finite interval (a, b). For convenience, it is shown as univariate and symmetrical. B represents a slice through a bivariate version of A, at right angles to the paper, to show the density gradient.

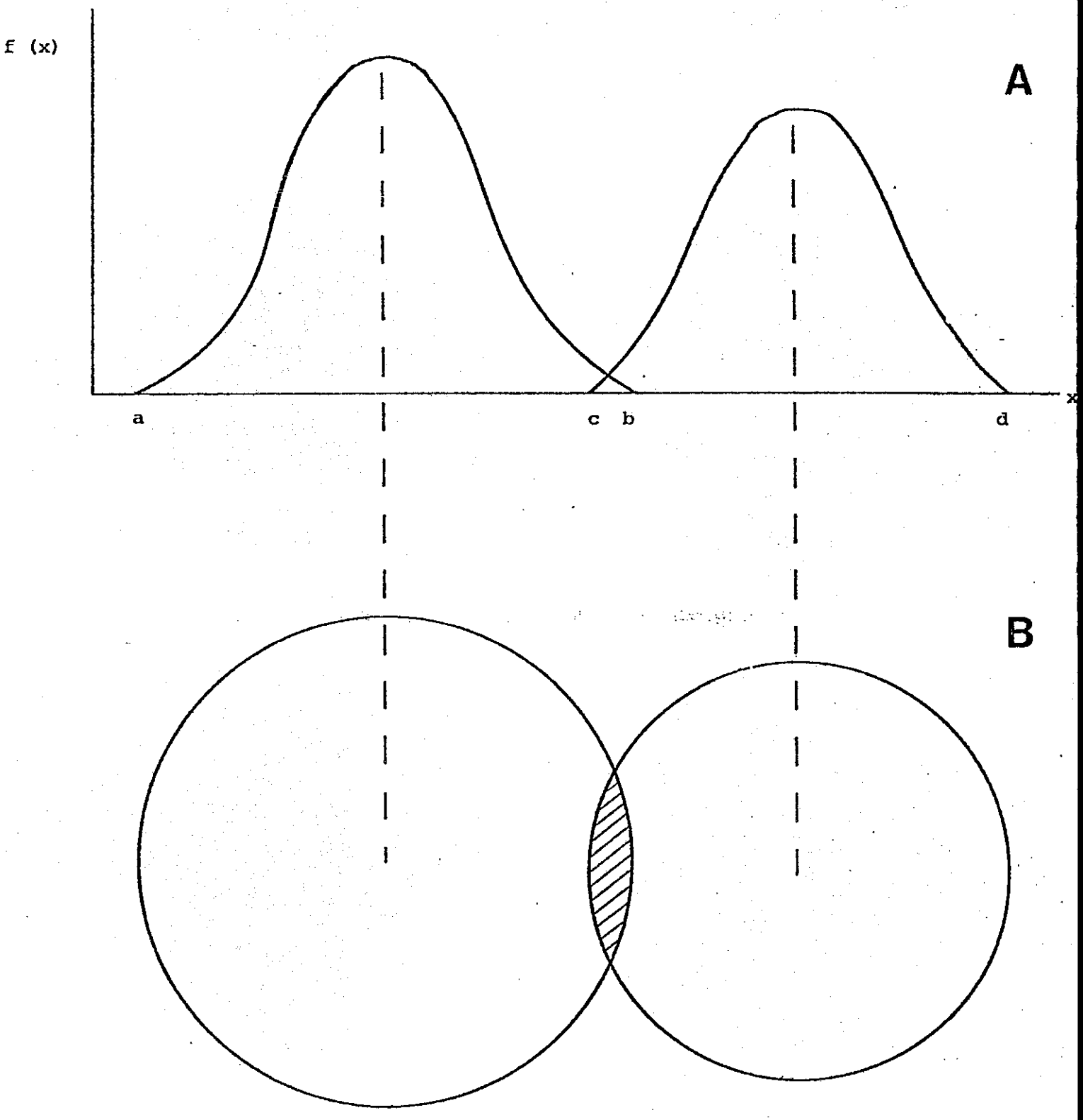


Figure 2. A bimodal sample with the tails of the frequency (probability density) distributions slightly overlapping.

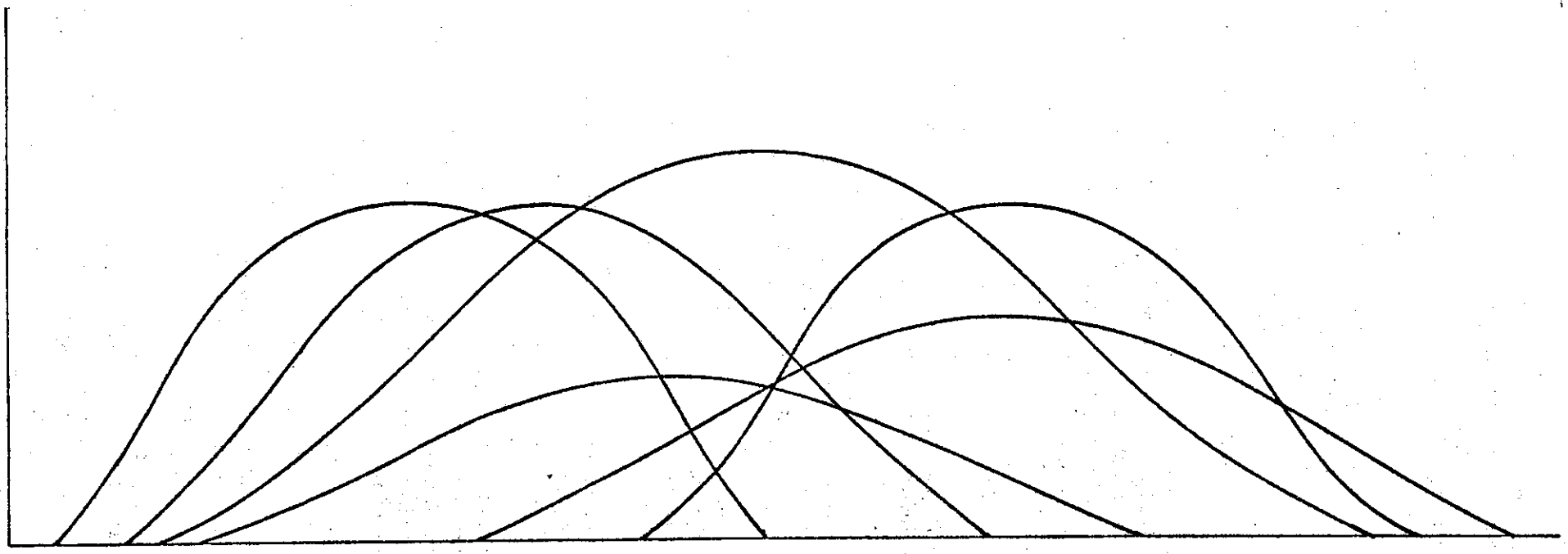

Figure 3. Complex overlapping distributions as visualized by some plant ecologists.

associations, and most difficult for the human brain to comprehend the data in totality. Where gradients are involved, ordination may be the best way of handling the data, although that, too, has its problems. The concept of such a unimodal continuum is not unique to ecology, examples can be found in other sciences (Clifford and Stephenson, 1975). The possibility that such a continuum may be divided into groups on the basis of so-called homogeneous areas led to the development of a variety of methods for this purpose, usually based on some sort of variance constraint. These methods, and the problems associated with them, will be discussed later.

## 2.5 Scaling, standardisation, and transformation of the data

In ecological studies, the raw data may not be uniform, because some species are more abundant than others for example, or because attributes measured on different scales differ in both range and variability. The importance of such differences in the analysis of the data needs thinking about, particularly in relation to the objectives of the investigator. In one context, it may be useful to exclude rare species (cf. Barkham, 1968). For some purposes, it may be judged that attributes should contribute equally regardless of their variation. On the other hand, the investigator may consider the variation itself to be an important feature to be retained.

We can recognize three ways in which the data may be modified: (1) Scaling; (2) Standardization; (3) Transformation. Scaling may be done in a variety of ways, the simplest being to add or subtract a constant from all values of a given attribute. Another method is to multiply or divide by a constant. Standardization means that the value of each attribute for each individual is expressed as a deviation from the mean of that attribute and divided by the standard deviation. This has the effect of reducing all attributes to unit standard deviation, and of reducing the magnitude of each attribute. Other methods are ranging (Gower, 1971) and rankits (Sokal and Rohlf, 1969). There appears to have been no employment so far in numerical taxonomy of standardizations that equalize the variability while leaving gross size unchanged (Sneath and Sokal, 1973).

The term 'transformation' is used of methods which seek to change the shape of the frequency distribution of the data, usually in the hope of obtaining an approximately normal distribution. In univariate statistics, for example, transformations may be used to satisfy the theoretical requirements of the analysis of variance. Classical multivariate theory has been based largely on the multivariate normal distribution, and although in some multivariate methods multivariate normality is not essential except for the sampling theory (e.g. in canonical correlation), not much is known about the robustness of the methods and the effects of large departures from normality. Some methods are sensitive to non-normality, for example cluster methods based on the assumption of a mixture of multivariate normal distributions. In numerical classification, dissimilarity measures may be sensitive to certain types of data. Thus, Euclidean distance types of measures, including variance measures, are particularly sensitive to data in which there are occasional very large values (Clifford and Stephenson, 1975).

Various types of transformation have been used (Sneath and Sokal, 1973; Clifford and Stephenson, 1975). Andrews et al (1971) discussed the problem of transformations in which the transformed variables are functions of the original variables collectively rather than separately, and suggested some techniques which might be useful. Clifford and Stephenson (1975) stated

that it remains uncertain whether the transformation required to produce normality of data is also the transformation which will produce optimal ecological 'sense', and that optimal ecological classificatory 'sense' is generally obtained by using a weaker transformation than that required to transform data to normality In any multivariate analysis, careful thought needs to be given to the nature of the data and how this relates to the methods to be used.

## 3. NUMERICAL METHODS

### 3.1 General

Two types of method have been found useful to describe the object-space; these are ordination and cluster analysis. There has been some discussion on the relative merits of these two types of method, as if they represented alternative ways of examining the data whereas they are, in fact, complementary. If there are natural groups, this should be apparent in the results of the ordination. If there are no such groups, ordination may still throw some light on the relationships between the individuals, and it is particularly useful if the individuals are distributed along gradients. Furthermore, ordination may show that a clustering method has been used for data to which it is not suited.

Many methods of cluster analysis break down for particular types of cluster that are, nevertheless, obvious to the eye. For example, elongated clusters are not well suited to Wishart's modal analysis, nor to methods that minimize $W^{-1}$ (see below); clusters with different dispersion matrices and outliers are not suited to methods which assume, explicitly or implicitly, homogeneity of within-group dispersions. Many methods tend to give clusters of approximately equal size, and failure to detect and eliminate an outlier can distort the procedure (Marriott, 1974).

Cluster analysis may involve searching for discontinuities in the object-space, or, more usually perhaps, the groupings are likely to be based on multimodality (Marriott, 1974). A totally different problem occurs if the data are unimodal; then the question is what is the best way of dividing the individuals into a given number of groups. This was called 'dissection' by Kendall and Stuart (1968, p. 314). It is important not to confuse this process with classification; there is no implication that the resulting groups represent in any sense a 'natural' division of the data, they are merely a matter of convenience and the only real criterion is their utility.

In both taxonomic and ecological classifications, it is possible to classify the entities by the attributes (previously called Q classifications, now called normal classifications) or the attributes by the entities (formerly called R classifications, now called inverse classifications). It is important that the models and principles on which the mathematics are based should be biologically well-founded and the conclusions reached tested against further experience. In the majority of cases, there are no absolute criteria against which to test the structure of a classification, and so it is important to be clear about the steps taken in its derivation.

### 3.2 Ordination

Initially, the positions of the entities in a multi-dimensional object-space are defined by their properties or some measure of their dissimilarity. Ordination procedures aim to preserve the relationships between the entities as accurately as possible and in a few dimensions. The reduction in dimensionality makes the data easier to handle mathematically: (1) it makes

graphical representation easier; (2) it removes difficulties which might arise from variables which are linearly related, or nearly so; (3) the variables resulting from the reduction may lend themselves to reification (i.e. the interpretation of the mathematics in terms of the original problem) and give a useful insight into the structure of the data (Marriott, 1974).

*Principal component analysis*

Possibly the best-known and most widely used ordination technique is principal component analysis (Anderson, 1958; Morrison, 1967; Seal, 1968; Blackith and Reyment, 1971). This involves a linear transformation of the attribute scores; the principal components are expressed in terms of linear combinations of the original variates.

The positions of the individuals can be plotted on pairs of right Cartesian component axes. Such plots will show discontinuities if they exist in the data (e.g. Blackith and Reyment, 1971), but it must be remembered that any such two-dimensional representation is distorted in that other dimensions are not taken into account. Gower and Ross (1969) showed how such distortions can be illustrated by superimposing the minimum spanning tree (see below) of th points in the total dimensionality on to their representation in the reduced space.

Holland (1969) pointed out that the vectors of the principal components are not the only ones capable of defining a space, and it is only a matter of geometrical manipulation to determine the extent to which the vectors of other components corresponding to biological hypotheses, or derived from other bodies of data, lie within such a space. Hence, it is possible to carry out the initial process a stage further and to transform principal components into other components which are either consistent with other results or more meaningful in the biological sense, giving a more general approach.

There has been much discussion about the use of principal component analysis and other ordination techniques in plant ecology; too many papers have been written for detailed discussion here. See, for example, Bray and Curtis (1957), Austin and Orloci (1966), Beals (1973), Noy-Meir (1973), Whittaker (1973), Orloci (1975).

*Principal co-ordinate analysis*

Principal component analysis is a special case of principal co-ordinate analysis, which operates on a matrix of some form of coefficient of association between all pairs of individuals (Blackith and Reyment, 1971). In principal component analysis, it is necessary to standardize the data unless all the variables are measured on the same scale. Hence, all attributes contribute equally to the total variance. The ability to ordinate a set of entities given only their dissimilarities can be useful in ecological studies, and there are some circumstances in which a particular dissimilarity measure might be preferred. For example, one might wish to emphasize dominance and thus use the Bray-Curtis measure, or, perhaps, be more concerned with relative properties and so use the Canberra metric (Clifford and Stephenson, 1975).

Principal co-ordinate analysis is particularly useful when there are missing values or missing variates. In such a case, a correlation type of similarity measure is reasonably robust and reliable, whereas replacing the missing values by estimates or guesses is not very satisfactory (Marriott, 1974).

*Factor analysis*

Blackith and Reyment (1971) stated that it is very hard to discuss factor analysis without generating more heat than light; it is the most controversial of the multivariate methods. Factor analysis was proposed originally as a model for a well-defined problem in educational psychology, but it acquired a bad reputation among mathematicians and was largely ignored outside the field of psychology, where it still finds most of its applications. The method and criticisms were discussed by Cattell (1965), Blackith and Reyment (1971) and Marriott (1974), and a book was written by Lawley and Maxwell (1971).

Factor axes may be rotated to determinable positions in which they are not necessarily, or even generally, orthogonal. Sneath and Sokal (1973) considered that this makes scientific sense in that the factors underlying the covariation pattern of the characters in nature are themselves undoubtedly correlated, but they pointed out that there are problems. Clifford and Stephenson (1975) went so far as to state "It is likely that in the future factor analysis will play an increasingly important role in ecological studies". On the other hand, Gower (1967b) considered it doubtful if factor analysis really is a helpful way of viewing biological data, and Blackith and Reyment (1971) asked: "Could it not be that factor analysis has persisted precisely because, to a considerable extent, it allows the experimenter to impose his preconceived ideas on the raw data?"

*Canonical variates and canonical correlation*

Allied to the above methods are two other multivariate techniques which investigate relationships in multi-dimensional space, but which operate on data which are already grouped either on the basis of individuals (canonical variate analysis) or variables (canonical correlation analysis). In canonical variate analysis, the relationships of the groups to each other in multi-dimensional space are investigated. As with the above procedures, the canonical variate space usually has a lower dimensionality than the original object-space. In canonical correlation analysis, the aim is to select pairs of maximally correlated linear functions from the two batteries of variables. Again, this reduces the dimensionality.

## 3.3 Cluster analysis

In cluster analysis, little or nothing is known about the category structure, all that is available is a collection of observations whose category memberships are unknown. The operational objective, therefore, is to discover a category structure which fits the observations. The partitions of the category structure should have various desirable properties (Jardine and Sibson, 1971; Anderberg, 1973; Sneath and Sokal, 1973; Clifford and Stephenson, 1975). In seeking structure in the data, two possibilities should be borne in mind: (a) the data may contain no clusters, i.e. the points are uniformly distributed in the measurement space and lack cohesion; (b) the data may contain only one cluster, i.e. there is a high mutual association among all points. Clearly, these two possibilities are extremes, with all other possibilities falling between them. Again, in searching for structure in the data, it should be borne in mind that any given set of data may admit of several different but meaningful classifications, each of which may pertain to a different aspect of the data. Furthermore, cluster analysis is a method for generating hypotheses. There is, as yet, no satisfactory definition of a cluster, and a classification obtained from a cluster analysis procedure has no inherent validity, its worth and its underlying explanatory structure is to be justified by its consistency with known facts. Cluster analysis methods involve a mixture

of imposing a structure on the data and revealing that structure which actually exists in the data. To a considerable extent, a set of clusters reflects the degree to which the data set conforms to the structural forms embedded in the clustering algorithm (Anderberg, 1973).

Jardine and Sibson (1971) pointed out that it has gradually been realized, in the last few years, that some of the variety of clustering algorithms which have been proposed, despite superficial differences, implement the same method, and that different methods differ very widely in their properties and results. They also stated that the development of a general theory of cluster analysis has been hindered by two widespread confusions. The first of these confusions is between algorithms and the methods which they implement. Thus, Lance and Williams (1967) have suggested as a general theory of hierarchic clustering what is, in fact, a generalized agglomerative algorithm, for the distinction is correctly applied to algorithms rather than methods. The second confusion concerns the role of models in data simplification. The term 'model' may be used in two quite different ways. One use covers the mathematical framework within which it is possible to analyse the properties of the methods of data simplification. The other use covers descriptions of algorithms in terms of their applications to some interpretations of the data. The latter may be called 'analogue models'.

Two kinds of analogue model have been widely used in cluster analysis. First, there are models which treat the objects as points or unit masses in Euclidean space (e.g. Gower, 1967a; Wishart, 1969a). Secondly, there are models which treat the objects as vertices of a graph, and values of the dissimilarity coefficient less than or equal to some threshold as edges (e.g. Estabrook, 1966; Jardine and Sibson, 1968a). Geometrical models can be applied only if the data are metric. Even when the data are naturally metric, a plausible geometrical interpretation for an algorithm does not necessarily provide any justification for the method which it implements. Thus, the various average-link and centroid algorithms which have simple geometrical interpretations suffer from very serious defects (Jardine and Sibson, 1971). The graph-theoretic models are more generally applicable, since any dissimilarity coefficient and any stratified clustering can be characterized by a sequence of graphs.

Before going on to consider in detail some of the more common clustering methods, we need to define some terms. A taxon is a taxonomic group of any nature or rank. Operational Taxonomic Units (OTU's) are the lowest ranking taxa employed in a given study; they may be individuals, averages representing species, exemplars of genera. For hierarchic cluster methods, the end-point of the process is a dendrogram, or tree-diagram in which numerical levels are associated with the branch points. The clusters specified at a particular level in a dendrogram have the property that they are pairwise disjoint, i.e. distinct clusters do not meet, and every OTU belongs to some cluster, possibly consisting of that OTU alone. A dendrogram is numerically stratified, i.e. it fulfils certain conditions (Jardine and Sibson, 1971). Non-hierarchic cluster methods may produce clusters which overlap, and the latter may be numerically stratified.

*Similarity and dissimilarity measures*

In talking of groups or clusters, we have the concept of nearness (similarity) of entities within a cluster, and of distance (dissimilarity) between entities in different clusters. Most cluster analysis methods start with some kind of similarity or dissimilarity measure, and a wide variety of

measures has been proposed (Jardine and Sibson, 1971; Sneath and Sokal, 1973; Clifford and Stephenson, 1975). Some reflect the need to accommodate particular forms of data, as, for example, those restricted to binary data. Others allow for unevenness in the frequencies of attributes, and minimize the influence of large or small values. Yet others are based on prior ideas concerning the statistical distributions of the properties measured. For a wide range of data, most of the indices are monotonic with respect to one another, but not all indices are interchangeable. Many of the indices have become neglected because they are mere variants of others, or because they have undesirable properties.

A dissimilarity measure is regarded as a 'metric' if it possesses four properties: (1) symmetry; (2) triangular inequality; (3) distinguishability of non-identicals; (4) indistinguishability of identicals. These properties are clearly useful, and measures which fail to satisfy any of these criteria should be regarded with caution. For example, Euclidean distance D is fully metric, but $D^2$ is not. It is worth noting that measures which are fully metric for complete data may become nonmetric if there are missing data.

The choice of a distance measure may be dictated by the nature of the data or by the special interests of the user. For the highly-skewed binary data obtained from presence/absence records of plant species, it would be usual to use an information statistic; the standardized Euclidean measure is unduly sensitive to the presence of rare species or the absence of common ones. On the other hand, for data with no strong outliers and no extreme skewness, Euclidean distance would be preferred. If the data were everywhere non-negative, with few zeros, but with occasional extreme outliers which the classifier does not wish to dominate the analysis, the Canberra metric is indicated (Williams, 1971).

As Euclidean distance depends on the scale of the variables, it is unlikely to have much meaning if some variables have a much greater range of values than others. Hence, it is generally used only when all the measurements are of the same type or when they have been standardized in some way (Marriott, 1974). Detailed discussions of the different measures can be found in Jardine and Sibson (1971), Sneath and Sokal (1973), and Clifford and Stephenson (1975).

*Stratified hierarchic cluster methods*

Most available stratified cluster methods are of hierarchic type, and the end point of the process is a dendrogram. Jardine and Sibson (1971) defined a dendrogram as a function involving a distance coefficient and satisfying certain conditions. One of the criteria for a distance coefficient to be a metric was the triangular inequality. If this criterion is relaxed so that:

$$d(A, C) \leq \max \left[ d(A, B), d(B, C) \right]$$

this condition is known as the ultrametric inequality, and distance coefficients satisfying it are called ultrametric. The ultrametric inequality insures that the pair-function implied by the dendrogram is monotonic. Lack of monotonicity is a serious defect.

In stratified hierarchic cluster methods, the smallest distance between distinct groups is found, and taken as the current level; all pairs between which this distance occurs are listed. The resultant graph, with groups as vertices and smallest distances corresponding to links, is divided into its connected components, and groups lying in the same connected

component are united to form a smaller number of larger groups. New inter-group dissimilarities are calculated in some way, and the process is repeated. The methods differ in the way in which the intergroup dissimilarities are calculated. The methods most commonly used are: (1) single link (nearest neighbour); (2) complete link (farthest neighbour); (3) unweighted pair-group using arithmetic averages, called by Lance and Williams (1967) the group-average method; (4) weighted pair-group using arithmetic averages; (5) unweighted pair-group centroid method, called by Lance and Williams (1967) the centroid technique; (6) weighted pair-group centroid method, called by Lance and Williams (1967) the median method. In pair-group methods, only one OTU or cluster may be admitted for membership at one time. This constraint may be relaxed to give variable-group methods.

In the single-link (nearest neighbour) method, new inter-group dissimilarities are not calculated. Instead, the original dissimilarities are retained, and clustering is based on the smallest distance from a point outside the group to a point inside the group. As a cluster expands, its outside members are nearer to the outside members of other clusters, and are thus more likely to link with them. Lance and Williams (1967) described this method as 'space contracting' and it is this property that is responsible for the well-known defect known as 'chaining', i.e. OTU's connected by intermediate OTU's are clustered together. However, it may well be that the chaining is simply an indication of the lack of any real discontinuities in the data. Certainly, the single-link method is conceptually and computationally very simple, and it has a large number of satisfactory mathematical properties. In particular, it does not suffer from discontinuity; Jardine and Sibson (1971) criticized alternative hierarchic methods (below) for their lack of continuity, which was regarded as being a far more severe defect than chaining in many applications. Jardine and Sibson (1971) proposed an axiomatic framework for cluster methods within which the single-link method is uniquely acceptable, and in that context its defects must be viewed as those of hierarchic classification itself. Sibson (1973) stated that since the defects of the single-link method are well-enough understood and of such a nature as to cause it to be misleading only rather rarely, the method itself should generally be acceptable.

Complete-link (farthest neighbour) clustering is the direct antithesis of the single-link technique. When two clusters join, their similarity is that existing between the farthest pair of members, one in each cluster. The method generally leads to tight, spheroidal, discrete clusters that join others only with difficulty and at relatively low overall similarity values. Lance and Williams (1967) called this method 'space-dilating', Sneath and Sokal (1973) listed it as monotonic, but Jardine and Sibson (1971) pointed out that in this method the output dissimilarity coefficients are not continuous functions of the inputs. The effects of this discontinuity are not predictable in practice, and can lead to completely misleading results.

To avoid the extremes of chaining on the one hand, and, on the other, small, tight, compact clusters that leave out many of the less easily affiliated OTU's, other clustering methods were developed. The average-linkage methods may be divided into the arithmetic average and the centroid methods.

Arithmetic average clustering computes the arithmetic average of the dissimilarity coefficients between an OTU candidate for admission and members of an extant cluster, or between the members of two clusters about to fuse. The arithmetic average may be unweighted, as in UPGMA (unweighted pair-group method using arithmetic averages) also called the unweighted average-link method or the unweighted pair-group method, in which each OTU

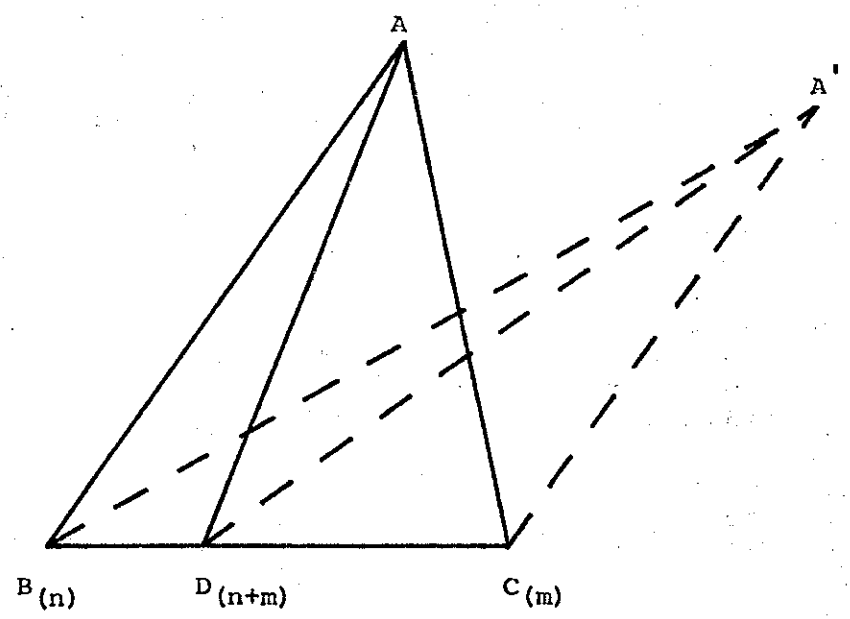


Figure 4. Centroid pair-group clustering. B and C are centroids of groups having n and m members respectively (n> m). BC is less than AB or AC. When B and C join, the centroid of the new group is D. In UPGMC, the ratio BD:DC is as m:n. In WPGMC, BD=DC. In either case, the point now representing B is nearer to A than was the original point, and the point representing C is farther away. The magnitude of this effect depends upon the relative positions of the points, it is more pronounced if A is in the position A'.

in a cluster is weighted equally. Or it may be weighted, as in the WPGMA (weighted pair-group method using arithmetic averages), also called the weighted average-link method or the weighted pair-group method. This differs from UPGMA by weighting the member most recently admitted to a cluster equal with all previous members, and distorts the overall taxonomic relationships in favour of the most recent arrival. Details of these methods are given by Gower (1967a), Sneath and Sokal (1973), and Jardine and Sibson (1971). The general taxonomic structure produced by UPGMA is similar to complete linkage analysis, but there are some fine distinctions. WPGMA shares the properties of UPGMA but distorts the overall taxonomic relationships in favour of the most recent arrival in a cluster. Sneath and Sokal (1973) listed both of these methods as monotonic, but Jardine and Sibson (1971) pointed out that the output dissimilarity coefficients are not continuous functions of the inputs.

Centroid clustering finds the centroid of the OTU's forming an extant cluster, and measures the dissimilarity (usually Euclidean distance) of any candidate OTU or cluster from this point. Centroid clustering has a simple geometrical interpretation, but no similar geometrical interpretation can be found for arithmetic average clustering. UPGMC (unweighted pair-group centroid method) weights each OTU in a cluster equally. When two clusters join, the resulting centroid is nearer to the centroid of the larger of the parent clusters (Fig. 4). Lance and Williams (1967) considered this method to be space conserving, but the ultrametric inequality requirement is not met and the outputs are not monotonic. The WPGMC (weighted pair-group centroid method) has been called the median method by Lance and Williams (1967) from its linear combinatorial formula first developed by Gower (1967a). This method weights the most recently admitted OTU in a cluster equally to the previous members. When two clusters join, the resulting centroid is midway between the centroids of the parent groups. This method shares the properties of UPGMC, including a lack of monotonicity. There are some differences in the resulting taxonomic structure caused by the heavier weight accorded to the late joiners of clusters (Sneath and Sokal, 1973).

Because of the lack of monotonicity, and consequent reversals in the dendograms, the centroid methods have been largely avoided, and the strategy may be regarded as obsolete (Clifford and Stephenson, 1975). There is a slight confusion in the literature over the use of the terms 'weighted' and 'unweighted'. The usage of Sokal and Michener (1958), used in this section, tends to be following in spite of the fact that it is the reverse of what might normally be used. This is because they fixed attention on the original individuals and not on the clusters which may have been derived from them (Gower, 1967a).

*Non-hierarchic cluster methods*

Non-hierarchic classifications are those that do not exhibit ranks in which subsidiary taxa become members of larger, more inclusive, taxa. The relative merits of hierarchic versus non-hierarchic classifications are difficult to evaluate. For traditional biological taxonomy, hierarchic classifications are required, and even in related fields it seems desirable to have higher ranking taxa that summarize common information about the majority of the members of the (polythetic) taxa. Non-hierarchic representation may be preferred when emphasis is placed on a faithful representation of the relationships among the OTU's rather than on a summarization of those relationships.

Jardine and Sibson (1971) discussed cluster methods which, by generalizing to allow stratified systems of overlapping clusters, succeed in avoiding

the defects of the methods outlined in the previous section, and also recover more information than does the single-link method, although this is achieved at the cost of greater complexity in the resulting classification. They proposed two sequences of non-hierarchic stratified cluster methods, Bk and Cu, which were shown to be satisfactory within Jardine and Sibson's axiomatic framework. Bk operates by restriction on the size of the permitted overlap between clusters, Cu operates by a restriction on the diameter of the permitted overlap between clusters, which is proportional to the level of the cluster. Bk is likely, in certain circumstances, to be more unstable under extension of range than is Cu. Whenever there are well-marked groups with intermediates, Cu is likely to produce clusterings which are more stable as the range is extended, because it is less vulnerable to alteration in the number of OTU's intermediate between clusters. Cu pays for this greater stability by requiring stronger assumptions about the significance of the underlying dissimilarity coefficient than does Bk.

*Other clustering methods*

A variety of clustering methods has been proposed. Most have not been widely taken up because they are mere variants of already established methods, because they have undesirable properties, or because they are impracticable in programming terms. These methods may be found in the text-books already cited.

Lance and Williams (1967) proposed a flexible clustering strategy, the characteristics of which could be changed by altering the value of a parameter. However, there is some danger in adjusting parameters until one obtains a result which is pleasing (but see below).

Edwards and Cavalli-Sforza (1965) suggested dividing the points into sets such that the sum of squares of distances between sets is a maximum. This defines what they mean by a cluster. Gower (1967a) drew attention to the collossal computational labour involved in the direct examination of all the possible partitions of N points. On a computer with 5 $\mu$ sec access time, it would take 100 hours for N = 21 and 54 000 years for N = 41. Orloci (1967) devised a criterion for overcoming this heavy computational load, but this is not monotonic, and reversals in its value can occur (Sneath and Sokal, 1973).

It has been pointed out by several workers (see Wishart, 1969b) that methods of this kind may divide dense clusters in an unacceptable manner. Gower (1967a) raised the question of whether the method should maximize intergroup sums of squares of the distances between group centroids. The sum of squares method takes into account the sample size of each cluster, and since some samples of equal importance in the overall classification in certain cases will be based on greatly unequal numbers of OTU's, the method based on maximizing distances between centroids may be preferable. However, the centroid method has the disadvantage that there may be points in one cluster which are nearer to the centroid of another cluster. With well-separated clusters, the maximum sums of squares and maximum distance between centroid methods will yield the same results, but so will most other methods. The real test of a method lies in its ability to deal with more challenging cases (Sneath and Sokal, 1973).

Various methods have been proposed which seek to minimize some function of the root mean square pairwise dissimilarity within elements of the partition and maximize the root mean square pairwise dissimilarity between members of different elements of the partition. They are sometimes called sum of squares or variance methods (Jardine and Sibson, 1971, Sneath and Sokal, 1973, Clifford and Stephenson, 1975). The properties of these methods are not well known, but Jardine and Sibson (1971) noted that there is not, in general, a unique partition on which the measure is optimized.

One example of such a method is that of Beale (1969), for which an algorithm was given by Sparks (1973). In this method, the user specifies the number of clusters required and the initial cluster centres. Initially, each observation is allocated to its closest cluster centre. The means of the clusters are then calculated and are taken to be the new cluster centres. The observations are then checked in turn to see if a move to a different cluster results in a decrease in the total sum of squares. Beale (1969) pointed out that this method may not find the best grouping (global optimum) but it does find one that could not be improved by moving any single observation to another cluster (local optimum). Sparks (1973) drew attention to the importance of the choice of initial cluster centres, although it is not clear how this choice is to be made. He also pointed out that the results obtained with different numbers of clusters are not necessarily hierarchic.

Friedman and Rubin (1967) proposed a method based on minimizing the generalized variance within the groups (i.e. the determinant of the pooled within-groups sums of squares and products matrix). This idea seems attractive at first sight, because it is equivalent to minimizing Wilks' criterion. However, it was criticized by Marriott (1974), who pointed out that if the data consist of samples from a mixture of unimodal distributions, the groups defined by this procedure will be the truncated centres of these dispersions mixed with the tails of other distributions. The dispersion matrix estimated within groups will not be an estimate of the dispersion matrix of the underlying distributions even if these are identical, and there is no reason to expect that it will be the same within the artificial groups found by the clustering process.

Marriott (1971) attempted to overcome the difficulties experienced by other workers using the generalized variance approach by assuming a uniform distribution as a null hypothesis. If the null hypothesis is true, the effect of an optimum subdivision on the generalized variance can be predicted. If subdivision of data into g groups reduces the generalized variance by much more than that predicted, it is reasonable to suppose that it corresponds to an inherent grouping in the data. The method appears to work reasonably well, although when the modes are near together and the distributions overlap considerably, separation may be impossible even for very large samples. On the other hand, some peculiar unimodal distributions of an extremely leptokurtic type may be subdivided. It is necessary to test each pair of groups in isolation to see whether they should be recombined. Advantages of this method are: (1) great flexibility in the data that can be handled and in the use of concomitant observations; (2) independence of scale and of linear transformations. Its disadvantages are: (1) the mathematical basis is not altogether solid, in particular the criterion for subdivision is rather arbitrary; (2) a significance test, though theoretically possible, does not yet exist; (3) the computational load is heavy (Marriott, 1974).

Some cluster methods assume that the individuals in the groups are multivariate normally distributed, and the problem is then one of separating mixtures of normal distributions. Beale (1969) noted that attempting to minimize the sum of squares of the deviations of the observations from their respective cluster centres is equivalent to maximum likelihood if all clusters are assumed to be (spheroidally) normally distributed with a common variance. Two methods based on a maximum-likelihood approach have been suggested.

Marriott (1974) considered the method of Day (1969) to be almost the only classification technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model (that the underlying distributions are multivariate normal with

equal dispersion matrices), investigates it by well-established statistical techniques, and provides a test of significance of the results. The fact that it is difficult to apply, and in many situations is unrealistic, reflects the complexity of the question that cluster analysis is trying to answer.

The other maximum-likelihood method is that of Scott and Symons (1971). Their approach was to make maximum-likelihood estimates of the means, variances and covariances, and the identifying parameters that assigned the sample points to the groups. The resulting estimates indicated that the identifying parameters should be chosen to minimize the generalized variance. Marriott (1974) considered their conclusions to be misleading, and could not justify the method of minimizing the generalized variance used in this way. However, he also pointed out (Marriott, 1975) that the assumption of underlying normal distributions with equal dispersion matrices is seldom strictly true in practice, and in many practical situations, when the proportions in the underlying distributions are approximately equal, minimizing the generalized variance gives a sensible and reasonably robust clustering procedure, although it is better regarded as a heuristic approach rather than an estimation process applied to a particular model.

Marriott (1971) noted that the smallest increase in the generalized variance occurs when an individual is added to the group for which Mahalanobis' generalized distance of the individual from the group mean is minimum. This suggests that once the cores of the groups are known, allocation by multiple discriminant or canonical variate analysis could be used to the same effect, provided that the theoretical requirements of these methods are satisfied.

Other cluster methods not involving a dissimilarity measure include association analysis and related methods. These methods were devised primarily for the classification of individuals described by binary discrete-state attributes. Williams and Lambert's well-known monothetic technique of association analysis (see e.g. Sneath and Sokal, 1973) divides a set of OTU's into two subgroups based on the two states of a single character chosen to maximize chi-square. The subsets are similarly divided and the process ends when a predetermined number of groups is reached or when the measure of homogeneity as expressed by chi-square has fallen below a critical level. Other authors (e.g. see Jardine and Sibson, 1971) determine the fit of each partition in terms of the information loss induced by the partition. Lance and Williams (1968) have adapted the association analysis method to the information statistic 2I. So-called polythetic analyses seek that bisection which minimizes the information loss or some other related function regardless of whether bisection corresponds to the range of the states of any of the selected binary attributes (Jardine and Sibson, 1971), see, for example, MacNaughton-Smith et al (1964) and MacNaughton-Snith (1965).

Methods of monothetic association analysis have been used in ecology, but Jardine and Sibson (1971) stated that the available methods are unsatisfactory in several respects: (1) they are ill-defined, data can be readily constructed for which no bisection induces a unique minimum information loss, (2) partition into two subsets at each stage is an arbitrary choice; (3) their application in taxonomy is restricted to discrete-state attributes which do not vary within populations. They further noted that the use of these methods appears to rest upon a confusion between classification and diagnosis. Monothetic association analysis produces a hierarchic classification by choosing a diagnostic

key based on the available attributes which is in a precise sense optimal, but the production of optimal diagnostic keys is not the primary purpose of classification in ecology or taxonomy.

*Adaptive methods*

Most clustering methods are non-adaptive, that is, the algorithm proceeds toward a solution by means of a fixed clustering method which may, to a greater or lesser extent, impose a structure on the data. However, an ideal clustering method would be adaptive. It would make an initial exploration of the data to find the types of clusters that are probably present, and would then modify the clustering algorithm to suit whatever structure is considered to be most likely. Some methods which attempt this were discussed by Sneath and Sokal (1973). Two in particular will be noted here, both based on the single-link method.

A number of clustering methods possess variance constraints, Wishart (1969b) discussed thirteen. Implicit in the minimum variance approach is the concept that clusters should have no significant overall variance or spread, and this implies that in the case of a unimodal swarm the distribution should be split into an arbitrary number of compact sections. Forgey (1964, 1965) argued that clusters should correspond to data modes, and there can only be as many classes as there are distinct modes. No variance constraint is implied, or should be induced, for when a mode is elongated rather than spherical, the distribution merely reflects some internal factor of variation for the corresponding class. Forgey interpreted a data mode as a continuous dense swarm of points separated from other modes by either empty space or a scattering of 'noise' data. The 'noise' data may result from sampling errors or they may be interpreted as those natural phenomena associated with the intersecting tails of disjoint continuous distributions. The cluster analysis problem is therefore to isolate the dense centres irrespective of the interference (Wishart, 1969a, b).

Wishart (1969a, b) took the single-link method as a basis for his 'mode analysis'. The satisfactory mathematical properties of this method have already been outlined. Wishart's solution to the clustering problem was to remove the 'noise' data, to cluster the remaining dense swarms by single linkage, and then to re-allocate each noise datum according to a similarity criterion. This was achieved by selecting a distance threshold r and a density limit k. From each OTU, the method tests whether k or more OTU's lie within r, if so, the OTU is considered 'dense' (this corresponds to counting the number of links to the OTU in a single-link clustering). The 'dense' points are then clustered by a single-link method at the threshold r, and the resultant clusters delimit the dense cluster nuclei. Each 'non-dense' point is then allocated to a cluster by some criterion. By fixing k and varying r a hierarchical classification is produced. A severe decision demand is placed on the user in selecting r and k.

Harriott (1971, 1974) concluded that the method of Wishart was the best available for detecting and identifying a natural grouping, and it is unlikely to produce a meaningless or misleading answer. However, he also pointed out that it is insensitive in detecting elongated modes, and the choice of the value of k may affect the conclusions. He (1971) made the following points: (1) The search for modes by dense points can lead to misleading results when continuous distributions are involved unless samples are very large; (2) The dense points are defined in terms of a 'spherical' scanning device. This has certain advantages; discrete distributions can be included and there is no problem of unwanted classification on the basis of a single variate. On the other hand, the method is scale-dependent, is rather sensitive to the inclusion

of highly-correlated variates, and the existence of genuine multimodality can be masked by the inclusion of irrelevant variates, especially if the modes are ellipsoidal rather than spherical.

A similar strategy was proposed by Shepherd and Willmott (1968). In this strategy, the data are clustered in two stages. The purpose of stage one is to determine which OTU's are most likely to be at or near the centres of groups. In this stage, a single-link method is used, followed by a process of discarding peripheral OTU's until only compact nuclei remain. The severity of this reduction process is determined by a group reduction criterion. This results in a series of cluster nuclei which are fed to stage two, in which the cluster nuclei are expanded using a modified pair-group average linkage method. A re-admission criterion determines how easy re-admission into a group should be.

These approaches retain the desirable properties of single-link clustering and overcome the problems caused by chaining, and ecologists are currently looking into their possibilities.

*The admissibility criteria of Fisher and van Ness*

Fisher and van Ness (1971) approached the problem of selecting a 'best' clustering procedure via decision theory, which tells us to restrict our attention to admissible decision rules. They listed nine admissibility conditions, and specified which of these were satisfied by the following five clustering methods (the number given with the method indicates the number of conditions it failed to satisfy): (a) nearest neighbour, 1; (b) furthest neighbour, 2; (c) minimum least squares - k fixed, 4 with one condition not applicable; (d) hill climb least squares - k fixed, 5 with two not applicable; (e) centroid, 5.

Here again, it is clear that the two single link methods satisfy the greater number of conditions. Fisher and van Ness did not fail these methods on monotonicity (cf Jardine and Sibson, 1971, failed the furthest neighbour method on this count), these methods failed only on the convex admissibility, which Fisher and van Ness admitted does not seem universal since it eliminates many reasonable clusterings.

Fisher and van Ness (1971) also noted that, if two admissible clustering schemes give different dendrograms, one might wonder whether the data were suitable for a tree structure. This would seem to be a legitimate use of the flexible clustering strategy of Lance and Williams (1967). Thoughtful use of the properties of different clustering methods can reveal certain properties of the data. For example, the nearest-neighbour method maximizes the minimum intercluster distance at each step. The furthest neighbour method minimizes the maximum cluster diameter at each step. If different dendrograms result from the use of the two methods, then both of the above objectives cannot be attained at the same time. A comparison of the two trees would be revealing. Since, at present, there is little knowledge of how to choose between many different methods of calculating similarity coefficients and hierarchical clustering algorithms, presentation of the data under various methods would give some, admittedly non-quantitative, information on reliability of any dendrogram obtained. A more concise method of presentation would be to run several methods and give the diameter of the set of dendrograms obtained. This would help avoid the computation time objection to Hartigan's (1967) approach, but not the need for a metric.

## 4. GRAPHS AND TREES

In the graph theoretical sense, a graph is a set of points (vertices) and of relations between pairs of vertices indicated by lines called edges. A set of entities and their dissimilarities may be represented by a graph, with the entities shown as vertices and the dissimilarity relationships between them shown as edges. In graph theory, the edges are not directly associated with a real value such as a dissimilarity or distance. However, it is possible to associate a real number with an edge, and this can be called its length. It is easy to see that with entities as vertices, the lengths can be dissimilarity measures. Relationship is indicated by the presence (or lack of relationship by the absence) of an edge between two vertices. Hence, by breaking the graph at various special levels based on the lengths of the edges, one can form clusters of vertices. The edges connecting a cluster of entities indicate the set of those entities that are more similar to each other than an arbitrary criterion.

The utility of the graph theoretical approach in this context is threefold. First, graphs serve as illustrative devices that enable many investigators to understand a variety of problems connected with cluster analysis. Second, the graph theoretical approach enables us to derive certain properties of clusters from well-established theorems of graph theory and also to employ graph-theoretical tools as solutions to specific problems. Third, they provide extra information when superimposed on ordinations (Sneath and Sokal, 1973).

This leads us to certain basic concepts in graph theory. A graph is said to be connected if every pair of distinct vertices is joined by at least one chain. A minimally connected graph contains only one direct or indirect path between every pair of vertices. Removal of one edge from such a graph disconnects it into two subgraphs, which are also maximal connected subgraphs because they have no proper supergraph which is connected. A graph is said to be a tree if it is connected and has no circuits. The removal of any one edge of a tree yields a disconnected graph, since the edge removed constituted the unique chain joining two vertices. Hence, a tree is a minimal connected graph. If all vertices of graph G are included in tree T, then T is said to span G. A minimum spanning tree has the smallest possible sum of the lengths of the vertices.

A special family of graphs is the family of directed graphs (also known as networks), which imply direction in the edges. A directed tree has edges with direction and a unique path from one vertex, called the root of the tree, to all other vertices. A conventional dendrogram is an example of such a graph.

Minimum spanning trees have been found useful as an additional perspective of taxonomic relationships in an ordination (e.g. Gower and Ross, 1969; Rohlf, 1970; Schnell, 1970). Some cluster analyses leave invariant the dissimilarities between certain pairs of objects. The set of elements left invariant by the single-linkage clustering method corresponds to the edges of the minimum spanning tree (Rohlf, 1974a). Gower and Ross (1969) drew attention to the value of the minimum spanning tree in single-linkage cluster analysis.

Wirth et al (1966) presented a computer method for cluster analysis based on graph theory (cf Estabrook, 1966). The method, essentially a form of single-linkage cluster analysis, is based on the partition of the collection of specimens (OTU's) into equivalence classes (clusters) which are maximal connected subgraphs. In using this method, Wirth et al gave to the vertices the values of the similarity coefficients between the pairs of OTU's. Clusters were described by the value of similarity associated with them, and the numerical expression of the isolation of a cluster (called its 'moat') was the value of the similarity coefficient at which the cluster would join with another cluster or OTU. The moat could be thought of as a measure of the empty space around a cluster. When this method was applied to 31 members of the Oncidiinae (Orchidaceae), the results were interesting, and the hierarchy showed good separation of clusters. The results were considered to be more satisfactory than those for other cluster methods with the same data. In some cases, clusters were linked by 'articulation points', i.e. specimens intermediate between two clusters. The graph-theory model provided a theoretical framework within which the nature of the relationships could be examined.

## 5. GENERAL CONCLUSIONS

It seems clear from what has been written so far that it is advisable to have a clearly-defined strategy for the application of numerical techniques to an ecological problem, in order to make it clear why the various steps have been taken, and to avoid the numerous pitfalls with which the subject is littered. Applying a particular technique to a set of data without a good reason may give misleading results, especially if the data do not happen to suit the method. Again, some methods imply that a particular structure exists in the data and, if it does not, the results can be misleading. For example, the user may think he has a realistic classification when in fact he may have an arbitrary dissection.

It is important to consider the nature of the data required. It seems clear that binary data are unsatisfactory and should be used only in certain restricted circumstances. There is also the possibility that the data may need scaling or transforming in some way. Unless the ecologist has a clear understanding of likely structure in the data, the next step will be an examination of the data to get some idea of its inherent structure and to look for discontinuities in multivariate space. Ordination techniques are useful for this purpose, and also reduce the dimensionality.

If discontinuities do occur, they should be obvious in this preliminary examination, and may provide a basis for classification (e.g. Blackith and Reyment, 1971). If clear discontinuities do not occur, this may be for one of two reasons: (1) there is only one cluster and all the entities belong to it; (2) clusters exist but 'moats' are obscured by 'noise' data or the fact that the frequency distributions of the clusters have overlapping tails. Here, the problem is to search for the presence of modes about which clusters can be formed. For this purpose, a simple single linkage cluster analysis appears to be suitable, and has the advantage that it can be used in conjunction with the minimum spanning tree, which provides a useful visual aid. We are currently testing this approach on two practical problems, and are also investigating the problem of allocating points to the cluster nuclei.

A hierarchical strategy optimizes a route between the individuals of which the sample is composed, via intermediate groupings, to a single group consisting of the entire sample. The groups through which the process passes are not necessarily optimal in themselves, the best route may be obtained at the expense of some slight reduction in homogeneity of individual groups. With non-hierarchical strategies, the structure of the individual groups is optimized, and the groups are made as homogeneous as possible. However, no route is defined between groups and their constituent individuals, or between groups and the complete sample (Williams, 1971).

Marriott (1974) stated that as a method of cluster analysis, if there is no special reason for imposing the nested structure of the dendogram, the strictly hierarchical methods have serious disadvantages. For example, to decide whether a division into two or three groups gives a better representation of the data, it is necessary to compare the best division into two with the best division into three, and hierarchical methods will not usually give both. Blackith and Reyment (1971, p. 277) stated that it seems likely that hierarchical techniques are almost always undesirable in theory, but the consequences of using hierarchical techniques when the structure of the experiment renders such a practice dubious, seem not to be very serious.

If discontinuities do not occur, or if they do not divide the data in a way which the researcher considers to be useful, then the problem is one of dissection, not classification. If dissection is to be carried out, the basis of the dissection must be clearly defined. For example, an ecologist may regard the vegetation as essentially continuously changing, but changing more rapidly in some regions than others. He will therefore wish to treat these zones of maximum gradient as if they were discontinuities, and to sharpen them by an appropriate technique (Williams, 1971). Some cluster methods have properties which make them useful for different types of dissection, for example the various minimum-variance methods and methods of the association analysis type. The flexible clustering strategy of Lance and Williams (1967) may also be useful for this purpose.

ACKNOWLEDGEMENTS

I am grateful to Mr. J. N. R. Jeffers for commenting on a draft of this paper, and to Mrs. D. M. Howard for drawing the figures.

REFERENCES

ANDERBERG, M. R. 1973. Cluster analysis for applications. New York and London: Academic Press. 359 pp.

ANDERSON, T. W. 1958. An Introduction to Multivariate Statistical Analysis. New York, London, Sydney: John Wiley and Sons, 374 pp.

ANDREWS, D. F., GNANADESIKAN, R., and WARNER, J. L. 1971. Transformations of multivariate data. Biometrics 27, 825-840.

AUSTIN, M. P. and ORLOCI, L. 1966. Geometric models in ecology II. An evaluation of some ordination techniques. J. Ecol. 54, 217-227.

BARKHAM, J. P. 1968. The ecology of the ground flora of some Cotswold beechwoods. Ph.D. Thesis, University of Birmingham.

BEALE, E. M. L. 1969. Euclidean cluster analysis. 37th Session of the Int. Statist. Inst. 99-101.

BEALS, E. W. 1973. Ordination: Mathematical elegance and ecological naivete. J. Ecol. 61, 23-35.

BLACKITH, R. E. and REYMENT, R. A. 1971. Multivariate Morphometrics. London and New York: Academic Press, 412 pp.

BRAY, J. R. and CURTIS, J. T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27, 325-349.

CATTELL, R. B. 1965. Factor analysis: An introduction to essentials I. The purpose and underlying models. Biometrics 21, 190-215.

CLIFFORD, H. T. and STEPHENSON, W. 1975. An Introduction to Numerical Classification. Academic Press. 229 pp.

DAY, N. E. 1969. Estimating the components of a mixture of normal distributions. Biometrika 56, 463-474.

EDWARDS, A. W. F. and CAVALLI-SFORZA, L. L. 1965. A method for cluster analysis. Biometrics 21, 362-375.

ESTABROOK, G. F. 1966. A mathematical model in graph theory for biological classification. J. theor. Biol. 12, 297-310.

FISHER, L. and van NESS, J. W. 1971. Admissible clustering procedures. Biometrika 58, 91-104.

FISHER, R. A. 1936. The use of multiple measures in taxonomic problems. Ann. Eugen. 8,376-386.

FORGEY, E. W. 1964. Evaluation of several methods for detecting sample mixtures for different N-dimensional populations. Amer. Psychol. Assn. Meetings, Los Angeles, California.

FORGEY, E. W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. AAAS-Biometric Soc. Meetings (WNAR), Riverside, California.

FRIEDMAN, H. P. and RUBIN, J. 1967. On some invariant criteria for grouping data. J. Amer. Statist. Assn. 62, 1159-1178.

GOWER, J. C. 1967a. A comparison of some methods of cluster analysis. Biometrics 23, 623-637.

GOWER, J. C. 1967b. Multivariate analysis and multidimensional geometry. Statistician 17, 13-28.

GOWER, J. C. 1969. The basis of numerical methods of classification. In "The Soil Ecosystem", ed. J. G. Sheals, Systematics Association Publication No. 8, 19-30.

GOWER, J. C. and ROSS, G. J. S. 1969. Minimum spanning trees and single linkage cluster analysis. Appl. Statist. 18, 54-64.

HARTIGAN, J. A. 1967. Representation of similarity matrices by trees. J. Amer. Statist. Assn. 62, 1140-1158.

HOLLAND, D. A. 1969. Component analysis: An aid to the interpretation of data. Expl. Agric. 5, 151-164.

JARDINE, N. and SIBSON, R. 1968. The construction of hierarchic and non-hierarchic classifications. Comput. J. 11, 117-184.

JARDINE, N. and SIBSON, R. 1971. Mathematical Taxonomy. John Wiley and Sons Ltd. 286 pp.

KENDALL, M. G. 1966. Discrimination and classification. In "Multivariate Analysis", ed. P. R. Krishnaiah. London and New York: Academic Press. 165-185.

KENDALL, M. G. and STUART, A. 1968. The Advanced Theory of Statistics. Vol. 3. London: Griffin and Co.

LANCE, G. N. and WILLIAMS, W. T. 1967. A general theory of classification sorting strategies. I. Hierarchical systems. Computer J. 9, 373-380.

LANCE, G. N. and WILLIAMS, W. T. 1968. Note on a new information-statistic classificatory program. Computer J. 11, 195.

LAWLEY, D. N. and MAXWELL, A. E. 1971. Factor Analysis as a Statistical Method. London: Butterworth's, 153 pp.

MACNAUGHTON-SMITH, P., WILLIAMS, W. T., DALE, M. P., and MOCKETT, L. G. 1964. Dissimilarity analysis: a new technique of hierarchical sub-division. Nature 202, 1034-1035.

MACNAUGHTON-SMITH, P. 1965. Some Statistical and other Numerical Techniques for Classifying Individuals. Home Office Res. Unit Rep., Publ. No. 6. London: H.M.S.O.

MARRIOTT, F. H. C. 1971. Practical problems in a method of cluster analysis. Biometrics 27, 501-514.

MARRIOTT, F. H. C. 1974. The Interpretation of Multiple Observations. London, New York, San Francisco: Academic Press, 117 pp.

MARRIOTT, F. H. C. 1975. Separating mixtures of normal distributions. Biometrics 31, 767-769.

MORRISON, D. F. 1967. Multivariate Statistical Methods. New York, London, Sydney: McGraw-Hill, 338 pp.

NOY-MEIR, I. 1973. Data transformations in ecological ordinations I. Some advantages of non-centering. J. Ecol. 61, 329-341.

ORLOCI, L. 1967. An agglomerative method for classification of plant communities. J. Ecol 55, 193-206.

ORLOCI, L. 1975. Multivariate Analysis in Vegetation Research. The Hague: Dr. W. Junk, 276 pp.

ROGERS, D. J. and FLEMING, H. 1964. A computer program for classifying plants II. A numerical handling of non-numerical data. Bio Science 14, 15-28.

ROHLF, F. J. 1970. Adaptive hierarchical clustering schemes. Systematic Zool. 19, 58-82.

ROHLF, F. J. 1974a. Graphs implied by the Jardine-Sibson overlapping clustering methods, Bk. J. amer. Statist. Assoc. 69, 705-710.

ROHLF, F. J. 1974b. A new approach to the computation of the Jardine-Sibson Bk clusters. Computer J. 18, 164-168.

SCHNELL, G. D. 1970. A phenetic study of the suborder Lari (Aves) I. Methods and results of principal components analyses. Systematic Zool. 19, 35-57.

SCOTT, A. J. and SYMONS, M. J. 1971. Clustering methods based on likelihood ratio criteria. Biometrics 27, 387-397.

SEAL, H. 1968. Multivariate Statistical Analysis for Biologists. London: Methuen and Co. 209 pp.

SHEPHERD, M. J. and WILLMOTT, A. J. 1968. Cluster analysis on the Atlas computer. Computer J. 11, 57-62.

SIBSON, R. 1973. SLINK: An optimally efficient algorithm for the single-link cluster method. Computer J. 16, 30-34.

SNEATH, P. H. A., and SOKAL, R. R. 1973. Numerical Taxonomy, The Principles and Practice of Numerical Classification. San Francisco: W. H. Freeman. 573 pp.

SOKAL, R. R. 1974. Classification: Purposes, principles, progress, prospects. Science 185, 1115-1123.

SOKAL, R. R. and MICHENER, C. D. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, 1409-1438.

SØRENSEN, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, (4), 1-34.

SPARKS, D. N. 1973. Euclidean cluster analysis. Appl. Statist. 22, 126-130.

WHITTAKER, R. H. 1970. Communities and Ecosystems. New York: Macmillan.

WHITTAKER, R. H. 1972. Evolution and measurement of species diversity. Taxon 21, 213-251.

WHITTAKER, R. H. (ed.) 1973. Ordination and Classification of Communities. The Hague: Dr. W. Junk.

WILLIAMS, W. T. 1971. Principles of clustering. Ann. Rev. Ecol. Syst. 2, 303-326.

WILLIAMS, W. T., LANCE, G. N., WEBB, L. J., and TRACEY, J. G. 1973. Studies in the numerical analysis of complex rain-forest communities VI. Models for the classification of quantitative data. J. Ecol. 61, 47-70.

WIRTH, M., ESTABROOK, G. F., and ROGERS, D. J. 1966. A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae). Systematic Zoology 15, 59-69.

WISHART, D. 1969a. Numerical classification method for deriving natural classes. Nature, 221, 97-98.

WISHART, D. 1969b. Mode analysis: A generalization of nearest neighbour which reduces chaining effects. In "Numerical Taxonomy", ed. A. J. Cole. Proceedings of the Colloquium in Numerical Taxonomy, Univ. of St. Andrews, Sept. 1968. 282-311. Academic Press.