ISSN 0308 3675

MERLEWOOD RESEARCH AND DEVELOPMENT PAPER
No 85

STATISTICAL CONSIDERATIONS IN LAND AND HABITAT CLASSIFICATION IN THE STUDY OF RABIES SPREAD

by
P J A HOWARD

Institute of Terrestrial Ecology
Merlewood Research Station
Grange-over-Sands
Cumbria
England
LA11 6JU

November 1981

Based on a paper prepared for the Workshop on Habitat Classification, Fox Populations, and Rabies Spread held at Keble College, Oxford, 15-19 September 1980.

CONTENTS

		Page
1	Introduction and background to the problem	1
2	Numerical methods which can be useful in habitat classification	3
3	Possible applications of classification and dissection techniques	6
4	Types of data	10
5	Data acquisition	14
6	Some published numerical classifications and dissections of habitats	16
7	Conclusions and future possibilities	19
	References	21

1 INTRODUCTION AND BACKGROUND TO THE PROBLEM

The front wave of the most recent European rabies epizootic has been spreading westwards at about 30 km/year since the 1940's, despite many attempts to halt, or eradicate, the disease. The data concerning this spread are subject to a large number of recording biases, and the few facts in which one can place confidence would seem to be: the rate of spread of the primary wave varies in different regions (30 to 60 km/year); the disease appears not to occur if the density of its main vector and victim, the Red fox *(Vulpes vulpes)* is below 0.2 foxes/km^2; the time between the passage of the front wave and the recurrence of secondary waves varies; case incidences appear to cluster spatially; the path of the epizootic is deflected by topographic features such as rivers and mountain ranges.

Simple temporal models of rabies in fox populations (Anderson *et al.*1981; Bacon & Macdonald 1980) show that the density of the host population is important in determining the time course of an epizootic. Heterogeneity in the spatial density of foxes implies that the spatial path will also be affected by the density of the vector. Thus, in order properly to determine the spatial dynamics of fox rabies, it is essential to know, at a minimum, the initial spatial density of foxes, ie. before the density is affected by the disease. The Red fox is extremely difficult to study in the field, and it is obviously not feasible to undertake a complete field survey of the whole of continental Europe, or even a large part of it, in order to determine fox population densities directly. However, fox population densities depend on characteristics of the habitat supporting them. If these characteristics could be identified by detailed local studies, and the relationship between habitat characteristics and fox population densities could be established, it should then be possible to estimate population densities in other habitats by recording the important characteristics.

Historically, it has been shown that there are differences in fox populations, and in patterns and rates of rabies spread, in widely differing habitats. However, the use of non-standard and subjective assessments of habitat type has precluded comparisons between results obtained in different areas or countries, or by different researchers. This limitation was recognized at a recent WHO symposium (WHO 1981). It has been suggested that numerical methods, eg. classification of fox habitats, could be helpful.

The purpose of this paper is to discuss theoretical and practical aspects of numerical classification procedures in relation to the study of rabies spread, and, in particular, to the problem of predicting fox (or other animal) population densities from environmental variables. Some other numerical methods which may prove useful are also mentioned. The aim is to suggest possibilities and to point out possible pitfalls. As our present understanding is not good enough for a prescription to be offered, some practical tests are necessary to compare the usefulness of the methods suggested here (see section 7).

We might usefully ask why is it thought that habitat classification can be of value in studying rabies spread? To answer that, we need to think about what we mean by classification, and what properties it is desirable for a classification to have in this context. Broadly speaking,

classification involves the recognition of similarities between, and the grouping of, objects or organisms. Classification is useful to describe the relationships of objects to each other, and to simplify those relationships so that general statements can be made about classes of objects.

An important distinction is between monothetic and polythetic classifications. The classes of a monothetic classification differ by at least one property which is uniform among the members of each class, so that possession of a unique set of features is both sufficient and necessary for membership of a class. A serious disadvantage of monothetic classifications for most ecological purposes is that they carry the risk of serious misclassification, since an object aberrant in the attribute used to make a division will be placed in a class away from objects which it otherwise resembles. Monothetic classifications are useful for certain special purposes, eg. in setting up taxonomic keys and in certain types of reference and filing systems. In polythetic classifications, on the other hand, groups of individuals or objects share a large proportion of their properties, but do not necessarily agree in any one property. Once a polythetic classification has been made, fewer properties are generally necessary to allocate objects to the classes than were necessary to establish the classes. Hence, classification of a data set results in a reduction of the amount of information that is necessary to describe the data set, but, if the classification is efficient, there is little or no reduction in the amount of information contained in the data.

For a fuller discussion of these general aspects, see Jardine and Sibson (1971), Sneath and Sokal (1973), Sokal (1974), Clifford and Stephenson (1975). From a statistical point of view, Kendall and Stuart (1968, p314) defined classification as the process of dividing a sample of objects, or an entire population, into groups which should be as distinct as possible. The groups should be 'natural' in the sense that members of any group should closely resemble each other and should differ considerably from those of another group. In practice, these criteria are usually interpreted by searching for discontinuities in the distribution, in multivariate space, of points representing the objects, or at least for regions of that space which are occupied by fewer points representing the tails of overlapping distributions and/or 'noise' data (see eg. Marriott 1974). By contrast, dissection implies the division of a sample or population into groups regardless of whether the boundaries are natural or not, the aim is simply to find the most convenient way of dividing the individuals into groups. Nevertheless, the groups formed by a dissection should have some definable logical structure.

All collections of objects can be dissected, not all can be classified. Classification may be a technique for generating hypotheses, but dissection is not, as the data are forced into a strait-jacket which restricts the domain of possible hypotheses and suggests that some will be generated by the process of dissection, rather than by the data (Cormack 1971).

As the objective is to predict population densities, it is reasonable to ask if regression methods might be more suitable. There have been few studies in which attempts have been made to predict animal population densities from environmental variables. Emanuelsson (1978, 1980) found that because many environmental variables are correlated, it was necessary to do a preliminary ordination by principal component analysis. The

resulting components were then used in regressions on known densities of particular bird species. Good predictions were obtained for a given area, but the regression for one area could not be used in a different geographical region. This suggests that a preliminary stratification of an area into more homogeneous regions is necessary, an approach which is discussed below.

When using numerical methods, it is important for the researcher to remember that the nature of the data has an important bearing on the numerical methods which can be used. It is important to think about this before the data are collected, and therefore this subject is also discussed in some detail.

2 NUMERICAL METHODS WHICH CAN BE USEFUL IN HABITAT CLASSIFICATION

In data analysis, two principles raised by Tukey (1954) should be borne in mind: (1) Different ends require different means and different logical structures; (2) While techniques are important ... knowing when to use them and why to use them is more important.

Two types of method which are useful in classification and dissection are ordination and cluster analysis.

Ordination

Ordination procedures aim to arrange points, representing objects, along new axes so as to preserve as much of the original information as possible, ie. to preserve the relationships between the objects as closely as possible. There are, ideally, fewer new axes than original variables. Ordination makes the data easier to handle mathematically in that: (i) it makes graphical representation easier; (ii) it removes difficulties which might arise from variables which are linearly related, or nearly so; (iii) the new axes may lend themselves to reification, ie. the interpretation of the mathematics in terms of the original problem, and so may give a useful insight into the structure of the data. If there are 'natural' groups, ie. groups which are separated by discontinuities in multivariate space or by regions of the space containing few points which represent the tails of overlapping distributions and/or 'noise' data, this fact should be apparent in the results of the ordination. If there are no such groups, ordination may still help to clarify the relationships between objects. Ordination may also be used to show if a clustering method has been applied to data for which it is not suited.

Possibly the best-known and most widely used ordination technique is principal component analysis (Anderson 1958; Morrison 1967; Seal 1968; Blackwith & Reyment 1971). This begins with a covariance matrix, or, more commonly, a correlation matrix, and the resulting components are expressed in terms of linear combinations of the original variables. Geometrically, the axes representing the variables are rotated to new positions (component axes) such that the first axis accounts for the maximum variance, the second axis accounts for the maximum possible variance in a direction orthogonal to the first, and so on. The rotations are orthogonal, ie. they preserve distances and angles, so that if the original variable-space is Euclidean, the component space will also be Euclidean. The most likely way in which the Euclidean properties of the variable-space will be lost will be that there are missing values in the data matrix.

Models in Euclidean space have at least three advantages (Williams & Dale 1965): (i) many simple, robust and powerful methods are available for dealing with Euclidean systems that are not available for non-Euclidean systems; (ii) they satisfy the requirement for hierarchical classification that each level in a dendrogram is associated with some measure which shall decrease as the hierarchy descends; (iii) it is easier to gain intuitive perception of Euclidean systems and to grasp their properties, and to predict those properties in extreme cases. Another useful property of Euclidean space is that the distance between any two points is unaltered by orthogonal rotation of the co-ordinate axes.

The positions of the objects can be plotted on pairs of rectangular Cartesian component axes. Such plots will show discontinuities if they exist in the data (eg. Blackith & Reyment 1971), but it must be remembered that any such two-dimensional representation is distorted in that other dimensions are not taken into account in the representation of inter-object distances. Gower and Ross (1969) showed how such distortion can be illustrated by superimposing the minimum spanning tree of points in the total number of dimensions on to their representation in the reduced space. There has been much discussion about the use of principal component analysis in plant ecology, where problems arise due to the patterns of distribution of plant species along environmental gradients (see eg. Noy-Meir & Austin 1970).

If the relationships among the objects are represented by an inter-object similarity matrix (different types of similarity measure are discussed by Sneath & Sokal 1973) then ordination by principal co-ordinates analysis (Gower 1966) can be used. Principal co-ordinates analysis is particularly useful when there are missing values or missing variates. In such cases, 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). An important feature of this method is that, as long as the similarity matrix has certain properties (see Gower 1966), the space defined by the principal co-ordinates axes is strictly Euclidean.

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 objects (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 fewer dimensions than the original variable-space. In canonical correlation analysis, the aim is to select pairs of maximally-correlated linear functions from two batteries of variables. Again, the dimensionality is reduced.

Cluster analysis

This term is applied to a wide range of techniques which seek to separate a collection of objects into groups or categories, there being little or no prior knowledge about the category structure of the data used in the analysis. To a greater or lesser extent, the different techniques involve the imposition of a structure on the data, as well as revealing any structure that may actually pre-exist. As a result, the groups that are identified reflect the degree to which the data conform to the structural forms inherent in the clustering algorithm (Anderberg 1973). Cluster analysis methods which have only a weak tendency to impose structure on the data, eg. single-linkage cluster analysis, are particularly

useful in exploratory data analysis.

A full discussion of clustering methods is outside the scope of this paper, see reviews by Cormack (1971) and Howard (1977). Clustering strategies have several important characteristics; they may be hierarchical or non-hierarchical, agglomerative or divisive, polythetic or monothetic. Hierarchical strategies find an optimum pathway between the objects of which a sample is composed, to a single group, consisting of the entire sample, via intermediate groupings. This pathway is found by a series of fusions (agglomerative) or, in the reverse direction, by a series of fissions (divisive), the groups produced being non-overlapping. The groups through which the process passes are not necessarily optimal in themselves, and the best pathway may be obtained at the expense of some slight reduction in homogeneity of the individual groups.

In non-hierarchical strategies, the structure of individual groups is optimized, and no pathway is defined between groups and their constituent individuals, or between groups and the complete sample. Marriott (1974) pointed out that a hierarchical strategy can have disadvantages if there is no special reason for requiring the nested structure of such a strategy. For example, if the aim is to decide whether a division into 2 groups gives a better representation of the data than a division into 3 groups, it is necessary to compare the best division into 2 groups with the best division into 3 groups, and a hierarchical strategy will not necessarily give both. Many of the most widely-used clustering algorithms employ agglomerative hierarchical strategies based on some sort of inter-object similarity or distance measure. Because the measure is based on several properties, such methods are polythetic. Agglomerative hierarchical strategies are inherently prone to a small amount of misclassification at the lowest, inter-object, level, where the possibility of error is greatest. On the other hand, with divisive techniques, there is a greater danger of inappropriate allocation of some objects that cannot later be corrected unless some special terminal reallocation procedure is used. Inappropriate allocation is particularly likely with monothetic techniques, because each division is based on 2 states of a single character, and any object which is aberrant in that character will be misclassified. Another problem with divisive techniques is that each group is made to divide into 2 at each level, an arbitrary restriction that may not reflect the inherent properties of the objects.

In searching for inter-object relationships which may be reflected in the data, the different possible patterns should be borne in mind. With most types of clustering algorithm, it is easy to identify the pattern in which distinct groups are separated by discontinuities in multivariate space, but this type of structure is by no means common. If the group centres are distinct, but the tails of the frequency distributions overlap, single-linkage cluster analysis will be unable to effect a clear division into clusters, although it can serve to indicate where the cluster centres might lie. Subsequently, other methods could be used to dissect the objects into groups, but, because the tails of the distributions overlap, a criterion is needed for the allocation of points in regions of overlap.

Where points representing the objects are more or less uniformly distributed in multivariate space and form a continuum, so that there is no clear structure in the data, classification in the strict sense is impossible; instead, the problem is overcome by dissection using criteria defined by the objectives of the analyst.

The shapes of clusters produced by cluster analysis alogorithms may also pose problems. In many clustering methods, some form of constraint is imposed on the spread of the clusters. Wishart (1969) discussed the properties of 13 such methods, and included them in the general category of 'minimum variance' methods. Some of the methods used by plant ecologists impose 'minimum variance' constraints. The minimum variance constraint makes these methods unhelpful, and even misleading, if the aim is to find the structure which actually exists in the data, unless it is known in advance that the structure is of a type for which the constraint is appropriate. However, methods which have this characteristic can be useful in dissection. On the other hand, single-linkage cluster analysis can identify clusters which are not only elongated, but also of complex shapes, if they are distinct.

In applying a hierarchical agglomerative clustering strategy, the user has a choice of similarity or distance measure and of clustering procedure. The choice needs to be made with some care. Similarity coefficients are normally appropriate to binary data and distance measures to continuous data, although some distance measures do have binary equivalents. Gower's (1971) general similarity coefficient can be used with mixed data types. The properties of the various measures are discussed in detail by Sneath and Sokal (1973) and Clifford and Stephenson (1975). Many of the commonly-used hierarchical clustering procedures have properties which, in at least some applications, are undesirable (Fisher & van Ness 1971; Jardine & Sibson 1971; Sneath & Sokal 1973).

This brief survey of numerical methods which can be useful in habitat classification suggests that a first step should be an ordination of the data. When the points representing the sampling locations are plotted on the ordination axes, relationships among the points can be examined. It should become clear whether or not classification *sensu strictu* is possible, or if a dissection is indicated. The researcher may choose to perform a cluster analysis on the ordination scores. Even if that is not done, the ordination charts are informative and useful. Which ordination procedure should be used depends upon the nature of the data. If all of the attributes are continuous, or if all of the attributes are binary with no more than, say, 30 attributes, and the data are not entirely of plant species presence or absence, then principal component analysis is appropriate. If the data are of mixed types, then principal co-ordinates analysis using Gower's general similarity coefficient should be used. The question of different types of data will be discussed in a later section. In the next section, the application of classification and dissection techniques will be discussed in more detail.

3 POSSIBLE APPLICATIONS OF CLASSIFICATION AND DISSECTION TECHNIQUES

There are two main ways in which classification or dissection could be useful: (1) as a stratification for future sampling, and (2) to enable properties to be predicted for new objects. We must then consider what properties a classification or dissection should have for these two purposes, so that we can choose an appropriate cluster analysis method.

Stratification

Stratification yields more efficient, and therefore more precise, estimators where the variables under consideration are homogeneous

within the strata produced but are heterogeneous in the overall population. The theoretical justification for stratification is in the reduction of sampling variances compared with simple random sampling. The more homogeneous the strata resulting from the stratification process, the larger will be the between-strata variance and the smaller the within-strata variance and sampling variance (Golder & Yeomans 1973). These criteria can be satisfied by the use of a k-means clustering algorithm. K-means clustering (Hartigan 1975) is a non-hierarchical method for producing a specified number of disjoint clusters such that the within-cluster sum of squares is minimized. No k-means algorithm produces a 'global optimum' solution unless N (the number of objects) is very small and there are only two groups. Instead, the aim is to produce a 'local optimum', ie. a solution for a given value of k (number of groups) such that no movement of an object from one cluster to another will reduce the within-cluster sum of squares. The algorithm given by Hartigan and Wong (1979) is very efficient. Such algorithms are called transfer (or iterative relocation) algorithms. In practice, it is often preferable to apply the k-means algorithm to component values after principal component analysis. The reason for this is that the method is strictly Euclidean; by using component values, problems of scaling of the variables and correlations among variables are overcome.

Some workers in I.T.E. have used indicator species analysis (ISA) as a method for stratification (Bunce *et al.* 1975). This method will not be discussed in detail here, as it has been discussed in detail in two other papers (Howard & Howard 1980, 1981). Ball and Williams used ISA on attributes from continuous data read from maps, for 436 10km x 10km National Grid squares with 'upland' characteristics. They chose 8 ISA classes for further study (I.T.E. 1978). Howard and Howard (1981) used principal component analysis and k-means clustering on the original continuous data. The k-means 8 groups obtained from the first 16 components of the product-moment correlation matrix and an overall sum of squares of 7334, that of the ISA 8-class partition was 9135. It was clear that ISA did not satisfy the minimum variance requirements for statistical stratification, and many of the ISA groups tended to be heterogeneous.

Principal component analysis cannot be used on disordered multistate or mixed-mode data. Ordered multistate data may not contain useful distance information. Gower's (1971) general similarity coefficient was designed for use with such data. Provided that there are no missing values, the resulting similarity matrix is positive semi-definite, which means that the similarities can be converted to distances with Euclidean properties. Application of principal co-ordinates analysis to the similarity matrix gives scatterplots of the points in a Euclidean space, and k-means clustering could be applied to those as for principal component analysis. An, alternative, or perhaps a complementary, approach (eg. see Aitchison 1978) would be to apply to the distance matrix a clustering technique with minimum variance properties (Wishart 1969).

While k-means clustering, based on data which are random variables, gives groups which have statistically desirable properties for stratification, groups so produced tell us nothing directly about the members of the groups, except that they should have similar properties. The within-group variation makes it impossible to make a precise prediction of the properties of a member of the group (unlike Gower's maximal predictive classification discussed below). However, a k-means stratification can be used as a basis for sampling, and the data collected in the sampling can then be

treated in a way which yields good predictions, eg. regression analysis. If we let d be the animal population density at a given location, then the value of d will be determined by some set of environmental factors.

We could write

$$d = f(e_1, e_2, ...)$$

where e_1, e_2, and so on represent environmental factors, and the relationship might be expected to vary with time.

This approach is illustrated by the work of Hirst (1975). In studying ungulate-habitat relationships in a south African woodland/savanna ecosystem, Hirst (1975) commented: "A natural community or group of communities can be regarded as a multivariate complex with the distribution of any specific organism therein being a function of the distribution of one or more biotic or physical community factors. Animals which exhibit a heterogeneous distribution over a given area are responding qualitatively and quantitatively to habitat factors which relate directly or indirectly to their well-being and survival. Certain of these factors may be so important that a relationship between them and the animal's distribution obviously exists. African woodland savanna is floristically, edaphically, and structurally complex, and animal-habitat relationships may not be easily discernible. Multiple regression using a digital computer offers a possible means of measuring the relative importance of a large number of habitat factors in collectively and individually determining the distribution of ungulates over a heterogeneous area".

Using a 300m x 300m grid marked on 1:20 000 scale aerial photographs, Hirst recorded the vegetation and used association analysis (Williams & Lambert 1959) to delineate 14 vegetational types or habitats. It is noteworthy that association analysis, now hardly used by plant ecologists, produces groups such that the distance between group centroids is maximal, ie. the groups tend to have minimum variance properties. A habitat map was prepared, and within each habitat the following features were recorded: Woody species - tree and shrub density, density of favoured browse species, shade cover, degree of clumping of woody plants; Herbaceous species - Herbaceous basal cover, grass and forb height, abundance parameters of favoured grasses and forbs, above-ground standing crop and production of herbaceous forage; Soils - physical composition, water infiltration into unsaturated soil in the wet season; Water-availability in pools. These features were recorded as 25 habitat characteristics.

The distribution of the various vegetational components was found to be largely determined by the same topoedaphic features, so that many of the vegetational characteristics measured were strongly correlated. A principal component analysis of the 25 habitat characteristics was carried out, and 10 components (accounting for 98% of the total variance) could be usefully identified in terms of ecological gradients. By sampling along transects, estimates of densities of each of 7 ungulate species were obtained. The estimated density within each habitat at any one time of sampling was taken as one observation of the dependent variable. There was thus a total of (no. of transects) x (no. of habitats) observations for each species. The independent variables in each case were the 10 principal components. Multiple regression, using second-degree polynomials where the relationships were non-linear, showed that each ungulate species had a unique combination of characteristics to which it responded in a positive or negative, linear or curvilinear, fashion.

Maximal predictive classification

As was noted above, while k-means clustering, based on data which are random variables, gives groups which have statistically desirable properties for stratification, groups so produced tell us nothing directly about the members of the groups, except that they should have similar properties. If, on the other hand, we had a number of groups whose attributes are preponderantly constant, then, because of high constancy and mutual interrelationships of attributes, such a grouping would carry a high predictive value for a new object assigned to it (Sneath & Sokal 1973 p188). When the sample objects are not subject to random variation, ie. when there is no variation of the selected attributes within groups, but the attributes change from group to group, the members of each group are completely identical, and any one member characterises the group. Repeated sampling will reproduce exactly the same sets of values. With k groups, a sample of objects will give us only k different sets of values however large the sample, the only information obtained by repeated sampling relates to the relative abundance of the different groups (Gower, 1970).

This situation arises naturally in the taxonomy of organisms, and leads to Gower's (1974) maximal predictive classification. If we have a matrix of v binary attributes for all the members of class C, then a binary row vector $\underline{m}$ of length v can be constructed which lists the properties that one would predict for an object on being informed that it belonged to that class. For the ith class, summed over all class members, there will be Wi correct predictions, and for all k classes there are $W = \sum_{i=1}^{k} Wi$ correct predictions.

The maximal predictive criterion selects that partition of n objects into k classes which maximises Wk. The average number Bk of properties correctly predicted for members of each class, using the class predictors of the other k-1 classes, measures the separation between classes. The best choice of k is related to maximising Wk - Bk. In practice, maximal predictive classification is implemented by using a transfer algorithm of the type used in k-means clustering, but using deviations from class predictors instead of from class means (Gower 1974).

In maximal predictive classes, all members of the ith class have more properties in common with their own class predictor $\underline{mi}$ than with the predictor of any other class. Therefore, an individual can be identified (ie. assigned to its correct class) by comparing it with the k class predictors. The sample belongs to the class giving most matches. Finally, it can be compared with members of that class until a perfect match is found.

With populations of inanimate objects, appropriate attributes may be easy to find. If suitable binary attributes cannot be found, multi-level qualitative data can be used (Gower 1974).

Gower's maximal predictive classification does not appear to have been much used in ecology. Curran and Swithinbank (1981) used it on a presence-absence data matrix for 110 quadrats and 110 plant species. The values for Wk-Bk suggested that the optimum number of classes was 7, and these classes represented developmental phases associated with management practices. Of the 110 plant species, 26 were used as class predictors, 11 of them appearing in more than one class.

It should be self-evident that classifications (or dissections) which are produced for the study of fox (or other animal) population densities are special classifications and require problem-specific information. A major problem for researchers is likely to be the selection of appropriate attributes. Because data can be recorded in different ways, and the type of data greatly influences the numerical methods which can be used, types of data are discussed in detail in the next section.

4 TYPES OF DATA

The fact that, in polythetic classifications, groups of objects share a large proportion of their properties but do not necessarily agree in any one property, has some important consequences. Firstly, it introduces the concept of similarity (or difference, or distance) among objects. In practice, this may be expressed in a variety of ways, some of which will be mentioned below. It also raises the problem of how many properties to use, and which properties to select.

As there is no theoretical basis on which to choose the number of properties used, the choice is likely to be made on practical grounds, and much depends on the objectives of the user. In ecological investigations, there is usually a finite but extremely large set of properties which could be recorded, but most of these are not recognized by the ecologist because they have no obvious practical relation to the topic being investigated. Thus, there is inevitably some degree of selection in any choice of properties to be examined.

Practical considerations are likely to limit the number of properties used. Apart from the fact that a large number of properties will result in more computer time, and storage space, being required in the analysis of the data, mathematical and statistical properties of the data matrix must be considered. The larger the number of properties chosen, the more likely it will be that many of the properties will be highly correlated, and will thus contribute no useful information, although they will contribute to the 'noise'. In some numerical methods, it is necessary to find the eigenvalues and eigenvectors of a matrix. In principal component analysis and factor analysis, the matrix has the order of the number of attributes. The larger the matrix, the less stable are the eigenvalues (eigenvalues and eigenvectors are important matrix properties, and are defined in basic matrix algebra text books such as Searle 1966). If there are too many binary attributes, values of similarity coefficients will depend on accidental matches or 'noise'.

Although it is important not to include too many attributes in a numerical analysis, the ecologist will be concerned that he may use too few, and thus omit an important attribute. An empirical approach to assist the user in deciding whether or not to add additional properties would be to perform an analysis on the original set of properties, add more, and repeat the analysis. If the results are similar, the classification is stable and the additional properties are not required.

A better approach would be to make a preliminary study of the proposed variables by principal component or factor analysis. In ecology, observed variables are often manifestations of a smaller number of factors of which the observer may be unaware. The variables which the investigator thinks are important may be informative, but they may not be the ones which are correlated with real structure (eg. see Muir 1962). Principal component or factor analysis can assist in the finding of important variables in the underlying factors.

Broadly speaking, a classification based on a large number of properties, chosen only because they are available or are easy to obtain, will be a general classification. It will serve a variety of purposes, but is unlikely to be optimum for any specific purpose. An example of a general classification is the group of plants that gardeners call "alpines", which share numerous growth and physiological characteristics reflecting their adaptation to alpine conditions (Sneath & Sokal 1973). On the other hand, if a classification is being constructed for some specific purpose, then there is a need for problem-specific information, and a classification based on selected properties is more likely to be optimal with respect to those properties, but might not be of general use.

The way in which the data are collected can have an important bearing on subsequent statistical and numerical analysis. Data consist of attribute scores. Conventionally, in statistics, the term 'attribute' is used for qualities possessed or not possessed, by an individual. The term 'variable' is usually used for quantities which may vary continuously. However, in pattern analysis the term 'attribute' has come to be used in a wider sense. This is convenient, since we do not need to differentiate between continuous and discrete data in general discussion where the nature of the data is not in question. There are many different kinds of attributes, see for example Sneath and Sokal (1973); Clifford and Stephenson (1975); Williams (1976). The most common kinds of attributes are: (i) Binary, eg. presence - absence; (ii) Disordered multistate (also called nominal attributes), such as colour or rock type; (iii) Ordered multistate (also called ordinal attributes), eg. rare, common, abundant; (iv) Meristic (discrete integer numbers), eg. number of petals; (v) Continuous, ie. measures on a continuous scale (also called quantitative or numeric attributes).

The scale on which the attributes are scored or measured will have an important influence on the subsequent data analysis. Presence-absence and disordered multistate data are on a nominal scale. With disordered multistate data, a given individual can be referred to only one state. The states may be numbered for computational convenience, but no meaning can be attached to the order in which the states are taken. An ordinal scale is used when various levels can be established for an attribute, but the scale values establish only the order of the observations, they do not contain any information on relative distances. With ordered multistate data, for example, rare, common, abundant, could be coded as 1, 2, 3, but these scores would not represent the relative abundances, ie the distances between the states are undefined. With interval and ratio scales, there is a concept of distance. On an interval scale, both the order and magnitude of an attribute state can be found relative to some arbitrary zero value, as in temperature scales. A ratio scale is used when the order and magnitude of an attribute state can be referred to some natural origin, as in measurement of length or weight. On such a

scale, the ratios between scale values are meaningful, as are the sums and differences. Probably few biologists have much formal background in types of measurement and scales, Torgerson's (1958) book is useful for this.

In numerical taxonomy, as in ordinary statistics, there is a greater choice of algorithms for dealing with continuous attributes than with binary attributes. Ordered and disordered multistate data can be a nuisance to handle, for many purposes there is no entirely satisfactory way of treating such data numerically. Meristic attributes, which can take only integral values, can also be a nuisance. In many cases, they cannot sensibly be treated as continuous; consider a case of counts of floral parts in a sample which contains some plants with two petals and some with four. As Williams (1976) pointed out, finding that the mean is three implies that dicotyledons have become monocotyledons. In such a case, the attribute could be coded as disordered multistate. On the other hand, in some cases the mean value is interpretable, and the data could be treated as continuous for practical purposes.

Continuous attributes are rarely continuous in the strict sense, since measurements are always made with limited accuracy, and there is always some degree of rounding off. In practice, the difference between continuous and discrete attributes depends on the chance that different observations take the same value. For practical purposes, counts that follow a Poisson distribution with a large mean can be regarded as continuous, since only a small proportion of the observations will take any one value. On the other hand, a continuous attribute grouped so coarsely that only a few values actually occur must, for at least some types of analysis, be treated differently. If an attribute takes a wide range of values, but has a concentration at one value, usually zero (eg. counts of parasites on a host), it may be better to score it as ordered multistate, eg. zero, low, medium, high (Marriott 1974).

Data may be coarsely grouped, either because measurements have been made with limited accuracy, or are replaced by rough assessments such as low, medium, high. For many purposes, such grouping is not important. The assessment can be replaced by suitable scores, giving whatever weight is considered appropriate to the differences between the groups. In many of the classical multivariate methods, the central limit theorem then justifies treating them as if they were jointly normally distributed. However, in some applications care is needed. This is especially true in cluster analysis. If the aim is to find a useful or meaningful grouping of the data, a coarsely-grouped attribute may exert a disproportionate influence on the result (Marriott 1974).

One way of dealing with disordered multistate data is to replace them by dummy binary attributes. Thus, the colours red, white, and blue could be coded as two binary attributes, one taking the value 1 for red and 0 for blue and white, the other taking the value 1 for white and 0 for red and blue. However, this method becomes rather cumbersome if there are many disordered multistate attributes, or many disordered states. This method can also be misleading if it is used in conjunction with an analysis that does not take into account the fact that the resulting binary attributes are correlated (eg. some forms of cluster analysis). With many observations of this type it is preferable to base a cluster analysis on some sort of similarity or dissimilarity measure (Marriott 1974). However, the comments of Rubin (1967), given below, should be taken into account.

Just as in univariate applications the standard error of a proportion can be used for significance tests and confidence intervals as if the proportion were normally distributed, so in the multivariate case the central limit theorem often justifies treating binary data as approximately multivariate normal. However, if all the data are of this type, other models are available, and some special methods have been evolved (Marriott 1974).

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 & Stephenson 1975, p30). In a botanical context, Greig-Smith (1964, p160) stated: "We are, in fact, dealing with a population of individuals (if stands may be so regarded) which differ from one another in terms of continuous variables of which presence and absence are only a crude expression". In plant and animal ecology, the tendency is to regard dominance or abundance (by some measure) as important (see eg. the papers in Section III of Ord *et al.* 1979). 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. Presence-absence data appear to work well when there are major differences in species distribution between sites, but they are not very useful for detailed studies of pattern if there are relatively few species with less clear-cut differences between sites. 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.

These attribute categories are not distinct, they depend to some extent on the sampling and coding procedure, and data in one form can be converted to another. In general, the conversion of continuous attributes (or discrete attributes that can be treated as continuous) to binary attributes is usually unsatisfactory. If a normally distributed variate is divided into two sections along the mean, all entities on either side of the mean would have identical binary scores, however far from the mean. Rubin (1967) drew attention to a difficulty which arises when a continuous attribute is chopped into a set of intervals each of which is scored as a separate attribute-state. He used the example of age, which could be changed to a discrete attribute of 8 states thus:

(1)	0 - 9	(5)	40 - 49
(2)	10 - 19	(6)	50 - 59
(3)	20 - 29	(7)	60 - 69
(4)	30 - 39	(8)	over 69

The obvious difficulty when using this approach is that two persons aged 29 and 30 will be regarded as dissimilar, whereas two persons of 30 and 39 will be regarded as similar. He suggested that in the calculation of similarities, the problem could be overcome by having the user specify two different criteria: (a) an interval, expressed in the units of a variate, such that two objects which have a difference on that variate smaller than the chosen interval will be considered to match (so that a one will be added to the number of matches when computing the similarity coefficient); (b) a second interval, expressed in the units of a variate, such that two objects which have a difference on this variate larger than the given interval will be considered not to match (and a zero will be

added to the number of matches). For differences which lie between these two specified values he suggests using linear interpolation. This method has the advantage of avoiding sharp discontinuities in a similarity coefficient when changes in the data are slight, and the burden on the user to provide values for the two intervals is no greater than that of breaking a continuous variate into arbitrary intervals. A similar method could be used for ordered multistate data.

Also, in the calculation of similarity coefficients, there can be problems with mixed attributes. Rubin (1967) drew attention to difficulties which occur when different attributes have different numbers of states. If we give equal weight to a binary attribute and to an attribute with four states, then, on average, the binary attribute will contribute matches more often to a similarity coefficient than will the four-state attribute. Furthermore, if we fragment the states of an attribute sufficiently, a match between two objects will become more and more rare, and the attribute will become useless in the analysis. Weighting attributes on the number of states is not the solution, since then an occasional match might completely dominate the similarity coefficient. The problem really lies in the original choice of states for each attribute. If there is too much variation in the number of states from attribute to attribute, then the many-state attributes will not play as important a part in a classification procedure as will the few-state attributes.

On the other hand, if one chooses states so that most objects assume only one or two of the states chosen, then one has thrown away information which could have increased our knowledge of the structure of the data set. Rubin suggested that the ideal solution might be to have an equal number of states for each attribute, and approximately equal frequencies for each state. This problem is not peculiar to the calculation of similarity coefficients. Once an attribute has been fragmented into too many states, or lumped into too few, information has been lost and is unrecoverable, no matter what the type of analysis. The particular problem for classification is that a randomly-chosen binary attribute may be more important in determining structure than several many-state attributes, even if the latter exhibit a high degree of structure when considered by themselves.

5 DATA ACQUISITION

Related to the above problems is the practical problem of how to acquire the data. This is really a subject which needs separate, detailed, treatment, and only some main aspects will be discussed here. In any land classification, data acquisition is complicated by problems of sample area. If data are obtained from maps, the scale of the map used is also important. The problem of the sample area is familiar to botanists in the problem of quadrat size. For example, if two species respond similarly to a controlling factor which has a defined pattern of values, they will show positive association up to the size of quadrat corresponding to the scale of heterogeneity of the controlling factor. Above that quadrat size, the indications of association will disappear (see eg. Greig-Smith 1964; Pielou 1969).

The problem of map scale is rather different. A map is only a pictorial representation of a portion of the earth's surface. With physical features, as represented on Ordnance Survey maps, the limitations of map scale make it necessary to simplify the representation of many surface features,

whilst other features, such as roads, may be exaggerated deliberately. In the production of maps, it is necessary to strike a balance between detail and clarity, and inevitably, some information is lost or distorted, according to the map scale (see Harley 1975).

Vegetation often plays a major part in land classifications. Dammon (1979) discussed vegetation properties which are useful in detecting and mapping vegetation patterns at various scales. His discussion of size of mapping unit for various map scales is worth reading by anyone interested in this topic.

Some information may be more accurately and easily obtained from air photographs, for example general slope angle and altitude. A single simple value expressing the altitude of a sample area could be obtained by taking the mean of N points located in the area. The problem with doing this on a map is that most of the points are likely to fall between contour lines, and it would be necessary to use non-linear interpolation to find the altitude of each point, which is not really practicable without complex equipment. On the other hand, it could be done easily using air photographs. Ball and Williams, in the study mentioned later, measured the proportion of land in different altitude bands in 10km x 10km National Grid squares on Ordnance Survey maps. There are two problems with this approach, firstly it is fragmenting a continuous property into intervals, as discussed above, and secondly, it requires several attributes to express one property. This means that the property, altitude, is effectively weighted, and could dominate an analysis irrespective of its ecological importance.

The estimation of slope angle from topographic maps by hand is time-consuming, and is liable to a considerable degree of subjectivity and approximation. Clerici (1980) described a method for the automatic drawing of slope maps from contour maps. His method is based on the determination of the slope (defined as the inclination of the plane which is tangential to the surface) at regularly-spaced points on a mathematical model of the topographic surface. First of all, altitudes and x-y co-ordinates of points on a topographic contour map are recorded using a digitizer. A computer is then used to superimpose a square grid on the set of data points and fit a polynomial trend surface to the area around each grid intersection. A denser grid is then superimposed on the computed trend surface, and the slope is calculated at each intersection. Finally, a map is produced by tracing the isolines obtained by interpolation of the slope points. The method can be developed to get further information such as the concavity or convexity of the surface and its aspect.

Yet another problem concerns the acquisition of data from 'factor maps', ie. maps of the distribution of special features or properties. MacDougall (1975) discussed sources of, and magnitudes of, error in factor maps. With soil maps, it must be noted that a soil mapping unit is a single expression of a multivariate system with a vector of means and a variance covariance matrix. If a property is deduced from a soil map, for any particular point, it is unlikely that any estimate of the likely accuracy of such a sample could be obtained. In the traditional approach to soil mapping, soils are identified in pits and the boundaries of mapping units are drawn by interpolation from auger borings using known relationships with landscape facets, geology, and vegetation. The mapping units are defined and described in terms of the soil series they contain. In most cases, one series dominates the mapping unit which then bears that series

name; more complex units carry the names of co-dominant series. In either case, the units contain lesser areas of other profile classes. The profile classes (soil series, variants, and phases) included in the mapping unit may be listed, and their frequency of occurrence assessed (eg. Clayden & Evans 1974). Various authors have discussed the concept of 'purity' of soil mapping units (eg. Bascomb & Jarvis 1976; Beckett & Bie 1975, 1976) as well as soil map accuracy (Legros 1973). The scale of mapping and the nature of the country impose what variation must be accepted (Ball 1964).

6 SOME PUBLISHED NUMERICAL CLASSIFICATIONS (AND DISSECTIONS) OF HABITATS

It is instructive to look at some published classifications (in this discussion it is convenient to include dissections and to use the term classification in its widest sense), to see what can be learnt from them. Thilenius (1972) investigated deer habitats in the Ponderosa pine forest of the Black Hills, South Dakota. In the published paper, his main effort was in the classification of the habitats, and little was done to relate the classes obtained to deer populations or activities. There is always a danger that classification may become an end in itself and, interesting though the subject may be, a classification or dissection should be shown to serve some useful end. Thilenius sampled the pine forest at 100 locations, each location being a "macro-plot" 60ft x 100ft. A total of 39 properties, giving 334 coded attributes, was recorded: Vegetation (3 properties) - frequency of overstorey trees, frequency of large shrubs, frequency of small shrubs, grasses, sedges and forbs, all in percent; Soils had 27 properties (recorded as mixed data types, continuous, meristic, disordered multistate and ordered multistate) as required by the standards of the U.S. Soil Survey; Site was represented by 9 properties of mixed data types.

The choice of properties is interesting. The vegetation types seem highly relevant to the problem. For some purposes, the physical structure of the vegetation cover may be more important than its species composition, for example in providing cover from predators and shelter from weather. The use of so many soil properties seems excessive. The choice of the standards of the U.S. Soil Survey suggests beaurocratic, rather than scientific, reasons for using so many properties, many of which would be unlikely to have much influence on deer populations. As with the vegetation cover, choice of soil properties should be relevant to the aims of the study. For some purposes, eg. in studying foxes, soil depth, stoniness, presence of indurated or compacted horizons, might be as important as soil chemical properties, since the physical properties are relevant to the construction of dens.

Thilenius standardized his diverse attributes by setting the maximum value for each attribute to 100. An inter-location similarity matrix was then calculated using a quantitative modification of Jaccard's coefficient (see Bray & Curtis 1957). It should be noted that in cluster analysis, the choice of a similarity measure needs careful thought, as different measures have different properties (see e.g. Sneath & Sokal 1973; Clifford & Stephenson 1975). Thilenius's similarity matrix was subjected to cluster analysis by the weighted pair-group centroid method (see Jardine & Sibson 1971; Sneath & Sokal 1973). This is an agglomerative hierarchical procedure which is now considered obsolete by many workers, as it has been shown to have some undesirable properties (Fisher & van Ness 1971; Jardine & Sibson 1971).

At the 0.54 similarity level, three clusters were obtained which, although they could be related in a general manner to the gross ecological and geographical features of the area, were highly variable with respect to other attributes and in the locations of which they were composed. At the 0.60 similarity level, 13 clusters were defined, and these produced a general ordering of the locations from the most xeric to the most mesic. The only attempt which Thilenius (1972) made to relate his habitat classification to deer populations and activities was to give a table of mean pellet group densities for the 13 clusters. A statistical test of these means suggested that the 13 clusters fell into 3 groups of "habitat units" which had similar mean pellet group densities. It is interesting that habitat units having similar pellet group densities belonged to more than one of the three clusters obtained at the 0.54 similarity level, which suggests that the numerical procedures used had not produced groups with good properties for defining the habitats with respect to deer use.

Radloff and Betters (1978) performed a somewhat similar classification of forest sites and for no stated purpose. They collected information on 147 square sites, each 2.4ha in area, located in a stratified random design within the Pike National Forest, Colorado. Six physical properties were recorded for each site from topographic maps: aspect (coded 1 for SSW to 14 for NNE); percent slope; altitude; terrain form (ridge top, spur ridge, level, swale - an old English word for a hollow or depression, valley bottom); local terrainrelief (straight, undulating, dissected); and position on slope (lower, middle, upper). Five soil characteristics were obtained from soil maps: minimum and maximum depths of the solum; structure; surface soil permeability; moisture capacity. Since the study has no particular aim, there is no basis on which to assess the value of their choice of properties.

Because of the mixed nature of the attributes, Radloff and Betters computed an inter-site similarity matrix using Gower's (1971) general similarity coefficient. This coefficient has two main advantages. First, it can be used with mixtures of binary, ordered or disordered multistate, or continuous attributes. Second, the similarity matrix is positive semi-definite unless there are missing values. This means that the N objects can be represented as points in Euclidean space (Gower 1966). The measure also gives the user the option of counting or not counting joint absences, ie. double zero matches.

Like Thilenius, Radloff and Betters used a centroid clustering procedure, which gave 13 clusters at the 0.8 similarity level. Canonical variates (multiple discriminant) analysis was then used to display the relationship among individuals and classes.

This study provides some interesting lessons. Radloff and Betters pointed out that their initial clustering gave groups which corresponded exactly with the soil series classification. The use of 5 correlated soil properties essentially resulted in soils-related information having a 5-fold weighting. To counteract this, each of the non-soil attributes was weighted 5-fold. Canonical variates ordination has been used by various workers to display the results of a clustering (eg. Grigal & Ohmann 1975). It can also be used to allocate new objects to existing groups. The method has the theoretical requirement that the groups should have homogeneous variance-covariance matrices, although there is a body of emphirical evidence available which suggests that the method may be moderately robust to departures from homogeneity. An alternative ordination would be by principal co-ordinates analysis (Gower 1966), using the Gower similarity coefficient. In practice, computing problems may be encountered with a

large number of objects, but Gower (1968) has shown how this can be overcome to some extent.

Omi *et al.* (1979) tackled the problem of combining similar land units within the Angeles National Forest, southern California, for the purpose of fire management planning, by a combination of three multivariate methods: (1) principal component analysis (to reduce the number of dimensions and provide orthogonal component values; (2) cluster analysis by an unweighted pair-group method using arithmetic averages (operating on Euclidean distances calculated from the component values); and (3) examination of, and reforming of, these initial groups by discriminant function analysis. The basic units (objects) of the analysis were major drainage basins which had been delineated as fire damage appraisal units in a previous study. The attributes used were those assumed to affect the long-term fire damage potential of a drainage basin, and they were chosen after a literature review and discussions with watershed scientists and managers in forestry, flood, and geological services. The data were collected from maps and reports, 10 characteristics were expressed in 15 variables.

Principal component analysis, followed by varimax (orthogonal) rotation, showed that the first 5 components accounted for 70% of the variation in the original 15 variables, and those components were used to calculate a Euclidean (presumably Pythagorean) inter-object distance matrix. No reason was given for preferring a hierarchical clustering procedure, or for the particular procedure chosen. One is often led to suspect that the choice of method is largely what happens to be in the available package. The clustering method used in this case has been criticised on mathematical grounds (Jardine and Sibson 1971), but in the context of the work of Omi *et al.* the method has the advantage of having minimum variance characteristics (Wishart 1969). A disadvantage is that the number of groups used has to be decided subjectively and Omi *et al.* selected 10 clusters. Discriminant function analysis reduced the number of clusters to 4. It would be very interesting to compare this result with the result of a k-means clustering.

Thompson *et al.* (1980) examined broad vegetation patterns of a land area of approximately 90 000km^2 in the Canadian Northwest Territories to determine the relative importance of areas of vegetation as Caribou habitats. The area was divided into 54 sampling units on the basis of LANDSAT data, and an aerial reconnaissance was made of sampling unit boundaries to ensure that sampling units which had the same theme pattern on the LANDSAT images appeared to have the same vegetation. The proportions of 8 previously-recognised vegetation cover types were found from sample transects in 43 of the 54 sampling units. Using the transect data, the 43 sampling units were submitted to cluster analysis using Ward's (1963) agglomerative hierarchical method, an iterative relocation procedure, and Wishart's (1969) mode analysis. Ward's method minimises the within-group sum of squares at each partition, and has minimum-variance characteristics (Wishart 1969). The three procedures were found to give similar results. Thompson *et al.* found, by discriminant function analysis, that 6 of the 8 vegetation cover types were significant in discriminating among the complexes. Analysis of pellet-group counts by cover types showed definite trends in seasonal use by Caribou.

On the North American continent, LANDSAT data are being increasingly used for mapping vegetation. Hathout (1980) described a technique, involving the use of a film density slicer with image enhancement, for using black and white LANDSAT transparencies to provide a vegetation map of Riding

Mountain National Park, Manitoba, Canada. This enhancement technique is said to be particularly useful in land use studies because the low resolution of space photographs does not lend itself to the direct extraction of relevant information. Enhanced images produced by this technique were found to provide reasonably accurate data for mapping the land cover of the National Park. Nine land cover classes were recognized, and were compared with a 1967 vegetation map of the area, as well as with field observations. The accuracy of the map produced by image enhancement was as follows:

Cover Class	Interpretation accuracy %
1. Lakes or swampland	92
2. Deciduous tree domination	84
3. Coniferous tree domination	87
4. Pure deciduous forest	73
5. Pure coniferous forest	87
6. Grassland or open forest	71
7. Very open areas (shrubland)	67
8. Burnt forest and marshland	86
9. Very dry areas (hilltops)	73

Hathout suggested that enhanced LANDSAT imagery might be used as a primary source for vegetation mapping, with very little assistance from ground survey of suspected areas of changes. It remains to be seen if this method is useful in Britain, where many changes occur in a relatively small distance.

7 CONCLUSIONS AND FUTURE POSSIBILITIES

There are clearly many problems which must be overcome if numerical methods, and particularly habitat classification, are to be used to predict fox population densities. The discussions in this paper suggest two broad approaches, using established numerical methods. In the first approach, a habitat classification (in the widest sense) is used as a basis for detailed sampling. From the data obtained in this sampling, the relationships between population density and environmental variables could be found by regression methods. In the second approach, fox population density is obtained directly from a maximal predictive classification.

1. The first approach requires an ordination of the data, the ordination space having Euclidean properties. With continuous environmental data, or with less than about 30 binary variables, the ordination could be by principal component analysis as long as the requirement for linear relationships among the variables is satisfied. With multistate, or mixed data types, ordination could be by principal co-ordinates analysis using Gower's general similarity coefficient. Both types of ordination give co-ordinates for the points, representing the sampling units, in Euclidean space (provided that there are no missing values).

Statistically, a stratification is required to have minimum within-strata variance and maximum between-strata variance. If ordination shows that there are clear discontinuities between groups of points, then a simple clustering procedure such as single linkage cluster analysis might be suitable. More usually, there will not be clear groupings and k-means clustering will be

necessary. A complementary exercise for mixed data would be to apply a minimum-variance hierarchical clustering technique to the similarity (or corresponding distance) matrix.

2. If the data are all binary, all multistate, or mixed types, it is possible to proceed directly to Gower's maximal predictive classification. Using this method, quantitative variables (eg population density) have to be treated as qualitative and the implicit ordering information has to be ignored. The way in which this is done will clearly affect the accuracy of the predicted fox population density.

Comparative tests of the possible approaches discussed above will be necessary before a researcher can be sure that a given method will satisfy the objectives of the study. The criteria in such tests will be the accuracy and precision of the predictions of fox population densities in relation to the cost and effort involved. Accuracy is defined as the closeness of predictions to the exact, true values. Precision refers to the dispersion of predictions from repeated observations about some centre, irrespective of whether or not the latter approximates to the true mean.

The real problem may well be to gather sufficient quantities of good data with which to establish the relationship between environmental factors and fox population density. In particular, problems of size of sample area need to be solved. The consequences of recording the data in different ways are discussed in detail in this paper, and the need for problem-specific information is emphasized. As yet, there seems to be no general consensus of opinion among fox ecologists as to what habitat factors are likely to be important. Indeed, there is evidence that in different types of habitat, different factors become important. More work is needed on these fundamental problems before good predictions can be made. A recent paper by Capen (1981) may be of some assistance in habitat response studies.

One practical problem in cluster analysis is what to do when the number of objects to be classified is unmanageably large. A solution suggested by Sneath (1964) is to run a random sample of the objects with the program, and from each well-defined cluster pick three objects as reference points for that cluster. Run a second sample including these reference objects, and repeat until all the objects have been processed. Many objects will belong to clusters previously recognised. The remaining 'solitary' objects should be re-run together with the reference objects to see if smaller clusters are formed. Using three reference objects per cluster also provides an internal check on the procedure, since they should always cluster closely together.

ACKNOWLEDGEMENTS

I am grateful to Dr. P.J. Bacon for helpful discussions and suggestions during the preparation of this paper.

REFERENCES

AITCHISON, J.W. 1978. Classification and mixed-mode data: an appreciation of Gower's general coefficient of similarity. Cambria, 6, 145-155.

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

ANDERSON, R.M., JACKSON, H.C., MAY, R.M. & SMITH, A.M. 1981. Population dynamics of Fox rabies in Europe. Nature, 289, 765-771.

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

BACON, P.J. & MACDONALD, D.W. 1980. To control rabies: vaccinate foxes. New Scientist, 87, 640-645.

BALL, D.F. 1964. Soil classification, land use, and productivity. Welsh Soils Discussion Group, Rep. No 5, 1-16.

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

BASCOMB, C.L. & JARVIS, M.G. 1976. Variability in three areas of the Denchworth soil map unit: I. Purity of the map unit and property variation within it. J. Soil Sci. 27, 420-437.

BECKETT, P.H.T. & BIE, S.W. 1975. Reconnaissance for soil survey I. Pre-survey estimates of the density of soil boundaries necessary to produce pure mapping units. J. Soil Sci. 26, 144-154.

BECKETT. P.H.T. & BIE, S.W. 1976. Reconnaissance for soil survey II. Pre-survey estimates of the intricacy of the soil pattern. J. Soil Sci. 27, 101-110.

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

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

BUNCE, R.G.H., MORRELL, S.K. & STEL, H.E. 1975. The application of multivariate analysis to regional survey. J. environ. Manage. 3, 151-165.

BURNHAM, K.P., ANDERSON, D.R. & LAAKE, J.L. 1980. Estimation of density from line transect sampling of biological populations. Wildl. Monogr. No 72, 202pp.

CAPEN, D.E. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA For. Serv., Gen. Tech. Rep., No RM 87, 104-113.

CLAYDEN, B. & EVANS, G.D. 1974. Soils in Dyfed I. Soil Survey Record No 20. Harpenden: Soil Survey of England and Wales. 148pp.

CLERICI, A. 1980. A method for drawing slope maps from contour maps by automatic data acquisition and processing. Comput. & Geosci., 6, 289-297.

CLIFFORD, H.T. & STEPHENSON, W. 1975. An Introduction to Numerical Classification. London and New York: Academic Press, 229pp.

CORMACK, R.M. 1971. A review of classification. Jl. R . Statist. Soc. Ser. A 134, 321-367.

CURRAN, P. & SWITHINBANK, P. 1981. The application of Gower's maximal predictive classification to vegetation data. J. Biogeogr., 8, 1-5.

DAMMAN, A.W.H. 1979. The role of vegetation analysis in land classification. For. Chron., 55, 175-182.

EMANUELSSON, U. 1978. A model for the prediction of bird communities based on vegetation data. Anser, suppl. 3, 80-83. (Swedish with English summary).

EMANUELSSON, U. 1980. Man and the bird fauna in the Torneträsk area. Fauna och Flora, No 1, 49-54. (Swedish with English summary).

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

GOLDER, P.A. & YEOMANS, K.A. 1973. The use of cluster analysis for stratification. Appl. Statist. 22, 213-219.

GOWER, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338.

GOWER, J.C. 1968. Adding a point to vector diagrams in multivariate analysis. Biometrika 55, 582-585.

GOWER, J.C. 1970. Classification and geology. Rev. int. Statist. Inst. 38, 35-41.

GOWER, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27, 857-871.

GOWER, J.C. 1974. Maximal predictive classification. Biometrics 30, 643-654.

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

GREIG-SMITH, P. 1964. Quantitative Plant Ecology. London: Butterworth, 2nd edition 256pp.

GRIGAL, D.F. & OHMANN, L.F. 1975. Classification, description, and dynamics of upland plant communities within a Minnesota wilderness area. Ecol. Monogr. 45, 389-407.

HARLEY, J.B. 1975. Ordnance Survey Maps, a Descriptive Manual. London: H.M.S.O. 200pp + 40 plates.

HARTIGAN, J.A. 1975. Clustering Algorithms. London, New York: Wiley, 351pp.

HARTIGAN, J.A. & WONG, M.A. 1979. Algorithm AS 136. A k-means clustering algorithm. Appl. Statist. 28, 100-108.

HATHOUT, S.A. 1980. Mapping vegetation by LANDSAT image enhancement. J. environ. Manage., 11, 111-117.

HIRST, S.M. 1975. Ungulate - habitat relationships in a South African woodland/savanna ecosystem. Wildl. Monogr. No 44, 60pp.

HOWARD, P.J.A. 1977. Numerical classification and cluster analysis in ecology: a review. Merlewood Research and Development Paper No 74, 29pp.

HOWARD, P.J.A. & HOWARD, D.M. 1980. Methods for classifying map data, with particular reference to indicator species analysis and k-means clustering. I.T.E. Ann. Rep. 1979, 34-42.

HOWARD, P.J.A. & HOWARD, D.M. 1981. Multivariate analysis of map data: a case study in classification and dissection. J. environ. Manage., 13, 23-40.

HUDSON, R.J. 1977. Habitat utilization and resource partitioning by wild ruminants: multivariate analysis of nominally-scaled attribute data. NW Sci. 51, 101-110.

I.T.E. 1978. Upland Land Use in England and Wales. Countryside Commission. CCP 111, 10-19.

JARDINE, N. & SIBSON, R. 1971. Mathematical Taxonomy. New York, London: Wiley, 286pp.

KENDALL, M.G. & STUART, A. 1968. The Advanced Theory of Statistics Vol. 3. London: Griffin, 557pp.

LEGROS, J.P. 1973. Soil map accuracy: The notion of fineness of characterization. Sci. Sol. 2, 115-128.

MACDOUGALL, E.B. 1975. The accuracy of map overlays. Landscape Planning 2, 23-30.

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

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

MUIR, J.W. 1962. The general principles of classification with reference to soils. J. Soil Sci. 13, 22-30.

NOY-MEIR, I. & AUSTIN, M.P. 1970. Principal component ordination and simulated vegetational data. Ecology 51, 551-552.

OMI, P.N., WENSEL, L.C. & MURPHY, J.L. 1979. An application of multivariate statistics to land-use planning: Classifying land units into homogeneous zones. For. Sci., 25, 399-414.

ORD, J.K., PATIL, G.P. & TAILLIE, C. (eds.) 1979. Statistical Distributions in Ecological Work. Statistical Ecology Series Vol. 4. Fairland, Md. USA: Int. Co-operative Publ. House. 464pp.

PIELOU, E.C. 1969. An Introduction to Mathematical Ecology. New York, London: Wiley, 286pp.

RADLOFF, D.L. & BETTERS, D.R. 1978. Multivariate analysis of physical site data for wildlife classification. Forest Sci. 24, 2-10.

RUBIN, J. 1967. Optimal classification into groups: an approach for solving the taxonomy problem. Jnl. theor. Biol. 15, 103-144.

SEAL, H. 1968. Multivariate Statistical Analysis for Biologists. London: Meuthuen, 209pp.

SEARLE, S.R. 1966. Matrix Algebra for the Biological Sciences New York, London: Wiley, 296pp.

SNEATH, P.H.A. 1964. New approaches to bacterial taxonomy: use of computers. A. Rev. Microbiol. 18, 335-346.

SNEATH, P.H.A. & SOKAL, R.R. 1973. Numerical Taxonomy. San Francisco: W.H. Freeman, 573pp.

SOKAL, R.R. 1974. Classification: purposes, principles, progress, prospects. Science, N.Y. 185, 1115-1123.

THILENIUS, J.F. 1972. Classification of deer habitat in the Ponderosa pine forest of the Black Hills, South Dakota. U.S.D.A. For. Serv., Res. Pap. RM-91, 28pp.

THOMPSON, D.C., KLASSEN, G.H. & CIHLAR, J. 1980. Caribou habitat mapping in the southern district of Keewatin, N.W.T.: An application of digital LANDSAT data. J. appl. Ecol. 17, 125-138.

TORGERSON, W.S. 1958. Theory and Methods of Scaling. London: Wiley, 460pp.

TUKEY, J.W. 1954. Unsolved problems of experimental statistics. J. Am. Statist. Ass. 49, 706-731.

WARD, J.H. 1963. Hierarchical grouping to optimize an objective function. J. Am. Statist. Ass. 58, 236-244.

WHO, 1981. WHO consultation on natural barriers of wildlife rabies in Europe. Geneva: WHO.

WILLIAMS, W.T. (ed) 1976. Pattern Analysis in Agricultural Science. Amsterdam, Oxford: Elsevier, 331pp.

WILLIAMS, W.T. & DALE, N.B. 1965. Fundamental problems in numerical taxonomy. Adv. bot. Res. 2, 35-68.

WILLIAMS, W.T. & LAMBERT, J.M. 1959. Multivariate methods in plant ecology I. Association-analysis in plant communities. J. Ecol. 47, 83-101.

WILLIAMS, W.T., LANCE, G.N., WEBB, L.J. & 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.

WISHART, D. 1969. Mode analysis: a generalization of nearest neighbour which reduces chaining effects. In "Numerical Taxonomy", ed. A.J. Cole, 282-311. London and New York: Academic Press.