Unresolved Problems in Cluster Analysis

B. S. Everitt

Biometrics, Vol. 35, No. 1, Perspectives in Biometry. (Mar., 1979), pp. 169-181.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197903%2935%3A1%3C169%3AUPICA%3E2.0.CO%3B2-Z

Biometrics is currently published by International Biometric Society.

BIOMETRICS 35, 169–181
March, 1979

Unresolved Problems in Cluster Analysis

B. S. EVERITT

Biometrics Unit, Institute of Psychiatry, University of London, SE5 8AF, England

Summary

The number of cluster analysis techniques has increased dramatically over the last ten to fifteen years, and they have been used in areas as distinct from one another as archaeology and psychiatry. However, a number of unresolved problems remain of which potential users of the methods need to be aware, if they are not to be faced with irrelevant or even misleading results. Such problems are discussed here, and an attempt is also made to indicate some possibilities for future work in the area.

1. Introduction

Classification in the widest sense is, along with astronomy, probably one of the oldest scientific pursuits undertaken by man. In the most general terms classification is the process of giving names to a collection of objects which are thought to be similar to each other in some respect. The ability to sort similar things into categories is obviously a primitive one, since it would seem to be a prerequisite of the development of language, which consists of words which help us to recognize and discuss the different types of events, objects and people we encounter; each noun in a language is a label used to describe a class of things which have striking features in common. Thus, for example, we name animals as cats, dogs, or horses and such a name collects individuals into groups.

Classification has played an important role in the development of many areas of science. Most notable, of course, has been its contribution to biology and zoology where it eventually led to Darwin's Theory of Evolution. It has, however, also played a central part in other fields. For example, the classification of the chemical elements in the periodic table, produced in its most complete form by Mendeleyev in the 1860's, has had a profound influence on the understanding of the structure of the atom. Again in astronomy the classification of stars into dwarf stars and giant stars using the Hertsprung-Russell plot of temperature against luminosity has strongly affected theories of stellar evolution.

In this paper classification is considered to be the process of allocating entities to initially undefined classes so that individuals in the same class are, in some sense, similar to one another. Techniques which generate classifications are variously known as *numerical taxonomy methods, methods for unsupervised pattern recognition*, and perhaps most commonly *cluster analysis methods*, and during the last two decades a vast variety of such techniques has been developed. The purpose of this paper is not to attempt a complete review of the area—such reviews are available elsewhere (see, for example, Cormack 1971 and Everitt 1974)—but to discuss some unresolved problems, and to speculate briefly on possible future develop-

Key Words: Cluster Analysis; Multivariate Analysis.

ments. We begin, however, with a short discussion comparing and contrasting cluster analysis with other multivariate analysis techniques.

2. *Cluster Analysis and Other Multivariate Techniques*

The raw data for many forms of cluster analysis is the familiar ($n \times p$) matrix of multivariate observations in which p variable values are given for each of n individuals. Such data is often also analysed by means of other multivariate techniques such as principal components and factor analysis, discriminant function analysis, multivariate analysis of variance, and increasingly by informal graphical methods. In this section we shall give a brief account of the relationship of cluster analysis to these other techniques.

2.1. *Cluster Analysis and Factor Analysis*

A distinction which is often made between these two sets of techniques is that cluster analysis is concerned with the classification of individuals, whilst factor analytic techniques assess relationships between variables and could be considered to be concerned with the classification of these variables. Such a distinction is however rather artificial since there is, essentially, no reason why many clustering techniques could not be used to cluster variables into groups, and it is also possible to use Q-mode factor analysis (see, for example, Cattell 1952) to directly classify individuals.

A more fundamental difference arises from consideration of the well formulated linear model of factor analysis. This has no equivalent in most methods of cluster analysis. Such a clearly defined model has advantages in leading to testable hypotheses concerning certain aspects of the structure of the data. It has disadvantages in respect to the linearity constraint, which the majority of users conveniently ignore. Such a model also has little meaning when applied to individuals rather than variables and consequently Q-mode factor analysis has been subjected to much criticism (see, for example, Zubin and Fleiss 1965, Fleiss and Zubin 1969 and Fleiss, Lawlor, Platman and Fieve 1971). The method has also been criticised on more pragmatic grounds, namely that it performs very poorly in practice, by Blashfield (1976).

In many respects the logical order of analysis on multivariate data should be a cluster analysis of individuals *followed* by separate factor analyses of variables *within* each group found in the previous stage. This might prevent factor analysis being carried out on data in which distinct groups were present and for which the overall correlation matrix would not necessarily be indicative of the within group correlations. However, such a process may in many cases not be possible simply because of the large number of variables involved in the raw data making it unsuitable for cluster analysis. In such cases the usual procedure is to perform the factor analysis first and then use a number of factor scores for each individual as input to the clustering method. As a necessary means of data reduction and simplification this seems acceptable; however a within group factor analysis following the clustering should still be performed, and these factors contrasted with those obtained initially.

2.2. *Cluster Analysis and Discriminant Analysis*

Discriminant function analysis is not a classification procedure *per se* since it requires an existing two (or more) group classification as starting point. However, it may often be usefully employed in association with cluster analysis as an informal indicator of which

variables have contributed most to cluster formation (the usual significance tests are not however valid; see next section), and as a means by which clusters and the relationships between them may be examined visually. Such canonical variate plots are used in the NORMAP program written by Wolfe (1970), and an example of their use in association with a cluster analysis is given in Everitt (1976).

2.3. Cluster Analysis and Analysis of Variance

A method of cluster analysis originally proposed by Friedman and Rubin (1967) uses Wilk's lambda statistic originally proposed in the context of multivariate analysis of variance, as its clustering criterion. This method will be subject of some discussion in Section 3 of this paper. Its connection with analysis of variance is mentioned here so that the problem of between clusters 'significance tests' can be discussed.

Wilks' lambda, $|T|/|W|$, arises from consideration of the fundamental equation

$$T = W + B \tag{1}$$

where T, W and B are $(p \times p)$ matrices containing, respectively, 'total', 'within' and 'between' sums of squares and products. The distribution of lambda is known under the null hypothesis that the g groups are samples from the same population, and this distribution is used in the analysis of variance context to assess the significance of differences between group mean vectors. Such significance tests are *not*, however, valid in cluster analysis applications since here we construct groups which *maximize* $|T|/|W|$, and it is, consequently, the distribution of $\max\{|T|/|W|\}$ under the null hypothesis, that we would need to study to answer questions of the statistical significance of the groups found by this form of clustering. Similar remarks hold for other clustering criterion, and for the significance tests sometimes carried out on individual variables after clustering using the usual t or F-tests. For example, Table 1 in Paykel and Rassaby (1978), contains a number of F-tests which are judged for significance using the usual tables of the F-distribution; such tests, if performed at all, should however be assessed against the critical values given in Englemann and Hartigan (1969).

Again in the analysis of variance context, cluster analysis has been proposed as an alternative to multiple comparison procedures for grouping means; see Scott and Knott (1974).

2.4. Cluster Analysis and Graphical Techniques

Recently there has been increased interest in the use of graphical techniques in statistics in general, and in multivariate analysis in particular (see, for example, Gnanadesikan 1977, Cox 1978 and Everitt 1978a). Graphical aids may simply be useful as a means of presenting the raw data, or they may, in some cases, themselves constitute the statistical analysis by providing both a summary of the informational content of the data and an exposure of unanticipated characteristics, such as possible inadequacies of the assumed model. Many of the proposed methods might usefully be employed in association with cluster analysis. For example, some low-dimensional representation of the data using one of the available ordination techniques might initially indicate whether applying some form of clustering is likely to be useful. Such plots can also be helpful in clarifying just how distinct are the clusters found. This might assist in preventing investigators claiming the discovery of some typology in their particular area when they have, essentially, merely split arbitrarily a homogeneous sample.

3. *Cluster Analysis—Some Unresolved Problems*

The number of problems associated with clustering techniques is legion. How should variables be scaled? Which distance or similarity measure should be used? How should clusters be tested for stability and validity? How should we assess the significance of clusters? Which method of clustering should we use? Here we shall discuss just a few of these problems and some of the more recent attempts to overcome them.

3.1. Hierarchical Techniques

We begin by considering the class of hierarchical clustering techniques. These are perhaps the most popular of all the multitude of cluster methods, and the literature surrounding them is enormous. The concept of the hierarchical representation of a data set was developed primarily in biology. The structures output from a hierarchical clustering method resembles the traditional hierarchical structure of linnean taxonomy with its graded sequence of ranks, with specimens grouped into *species* and these groups themselves grouped into *genera*, etc. Although any numerical taxonomic exercise with biological data need not replicate the structure of traditional classification, there nevertheless remains a strong tendency among biologists to prefer hierarchical classifications. However, these methods are now used in many other fields in which hierarchical structures may not be the most appropriate, and the logic of their use in such areas needs careful evaluation. For example, in their biological applications questions concerning the optimal number of groups do not arise—here the investigator is specifically interested in the complete tree structure. Such questions are however raised by other users of these techniques, who consequently require a decision regarding that stage of the hierarchical clustering process which may be regarded as optimal in this sense. Informal methods which have been suggested for this purpose are generally of the type where the dendrogram is examined for large changes of level, this being taken as indicative of the correct number of groups. However, Everitt (1974) shows that such a procedure may in many cases be misleading; it appears that a large change in fusion level in a dendrogram is a *necessary* but not a *sufficient* condition for the presence of clear-cut clusters. A slightly more formal approach to the problem is taken by Mojena (1977) who describes two possible 'stopping rules'. From empirical studies described in the paper, one of these rules does appear worthy of further consideration as a pragmatic means of objectively assessing the selection of a particular partition from a hierarchic clustering.

The late 1960's saw the first attempts at constructing a theoretical framework within which to study the properties of hierarchical techniques. Johnson (1967) showed that hierarchical clusters correspond to a distance metric which satisfies the *ultrametric inequality*, and that consequently a hierarchic dendrogram is characterised by an ultrametric. Since the input similarities or distances are not generally ultrametric (and only occasionally metric), Jardine and Sibson (1968) suggest that a cluster method which transforms a similarity matrix into a hierarchic dendrogram should therefore be regarded as a method whereby the ultrametric inequality is imposed on a similarity coefficient. They then specify a number of criteria which they argue it is reasonable for any such transformation to satisfy, and prove that single-linkage is the only method satisfying all the criteria, the implication seemingly being that it is therefore the only acceptable method. This conclusion has led to a certain amount of controversy. For example, Williams, Lance, Dale and Clifford (1971) question the need for cluster methods to satisfy *all* of Jardine and Sibson's proposed criteria, and adopt a more pragmatic approach to clustering, insisting that in practice single-linkage did not provide solutions which investigators found useful. Again, Gower (1975) feels that Jardine and Sibson's rejection of all but single-linkage clustering is too extreme, and questions

whether their criteria are not too stringent. His conclusion is that some of the criteria are *not* essential. It must be said that the approach taken by Jardine and Sibson appears to have had little impact on the majority of cluster analysis users; single-linkage is not particularly popular and the alternative mathematically acceptable method provided by these two authors is applicable only to small data sets and the solutions given are generally extremely difficult to interpret.

An alternative and very promising approach to understanding and evaluating the variety of hierarchic techniques available is to compare the effectiveness of different methods across a variety of data sets generated to have a particular structure. In this way the solutions obtained by a particular technique may be compared with the generated structure. Several studies of this type have been undertaken (for example, Cunningham and Ogilvie 1972, Kuiper and Fisher 1975, and Blashfield 1976). In general the results of such studies indicate that (1) no single method is best in every situation (2) the mathematically respectable single linkage is, in most cases, the *least* successful for the data used and (3) group average clustering and a method due to Ward (Ward 1963), do fairly well overall. Such empirical studies can, of course, never afford a complete evaluation of clustering methods; the results obtained do however, appear to indicate that Williams, Lance and co-workers are correct in the pragmatic approach they take and that there are more *useful* clustering methods than the mathematically acceptable single linkage technique.

On the other hand the single linkage method does have a number of desirable properties, perhaps the most important of which is that its results are invariant under monotonic transformations of the similarity matrix. (Other monotone invariant methods have been suggested by Hubert 1973 and D'Andrade 1978). This has led various authors to adapt the method in some way so as to retain its useful mathematical properties but to make it more practicably relevant. Examples are the methods proposed by Wishart (1969) and Zahn (1971). In addition Sibson (1973) has produced a very efficient algorithm for the technique which enables it to handle very large data sets and this may be regarded as a distinct advantage in many practical situations.

3.2. Clustering by Optimizing a Predefined Measure

Let us now move on to consider those clustering techniques which seek a partition of the data into k groups by attempting to optimize some predefined numerical measure indicative of a desirable clustering solution. Such methods differ from the methods discussed above in that the solution does not portray hierarchical relationships among the entities. The clusters denoted in a partitioning solution are discrete and exist at a single rank. For the moment we shall assume that the value of k is given *a priori*; the problem of deciding on an appropriate value for k will be discussed in detail later.

Several numerical criterion have been proposed for this approach to clustering. The most common is minimization of trace (W), a criterion which has been discussed by Friedman and Rubin (1967), McRae (1971) and Gordon and Henderson (1977). According to a survey of the published uses of classification in 1973 conducted by Blashfield (1976), this method is, in fact, one of the three most popular techniques of cluster analysis. It does however suffer from a number of problems. Firstly the method is transformation dependent; in general different results will be obtained from applying the technique to, say, the raw data, or to the data standardized in the usual way, that is to zero mean and unit standard deviation. This is of considerable practical importance in many applications where variables are on different metrics and some form of standardization is, in general, unavoidable. A further problem with the min{trace (W)} criterion is that the clusters produced are constrained to being hyper-

spherical; in cases where the real clusters in the data are of some other shape this may produce misleading solutions. Examples are given in Wishart (1969) and Everitt (1974).

The transformation dependency problem of the min{trace (W)} criterion led Friedman and Rubin (1967) to suggest other numerical cluster measures invariant to non-singular linear transformations of the data. Amongst these the one that has become most popular is minimization of det (W). Friedman and Rubin were led to this criterion by consideration of Wilks' lambda used as a test statistic in multivariate analysis of variance. Scott and Symon (1971) show how it arises using likelihood ratio considerations and Binder (1978) using a Bayesian approach to clustering shows it may be justified as maximizing certain approximated posterior probabilities. Apart from its advantages with regard to standardization considerations it has a further point in its favour, namely that it does *not* restrict clusters to being hyperspherical. It does however assume that all clusters in the data have the *same* shape, and again this can be a problem when the actual structure is not consistent with this requirement; see Everitt (1974) for an example. Some suggestions for overcoming this particular disadvantage of the det (W) criterion are made by Scott and Symon (1971), and Maronna and Jacovkis (1974). The former authors suggest as a clustering criterion

$$\min \left\{ \prod_{i=1}^{k} |W_i|^{n_i} \right\}$$

where W_i is the within group scatter matrix of group i, which contains n_i individuals. (The restriction that at least $p + 1$ observations must be assigned to each group avoids the degenerate case of infinite likelihood). An illustration of how this criterion performs more successfully than the simpler det (W) alternative when the clusters do have different shapes is given in Everitt (1974). Maronna and Jacovkis, in an interesting discussion of the metrics used in cluster analysis, suggest the criterion

$$\min \left\{ p \sum_{i} (n_i - 1) |W_i|^{1/p} \right\},$$

but this does not appear to have yet been used in practice.

3.3. Optimization Algorithms

Once a suitable numerical clustering criterion has been devised, consideration needs to be given as to how to choose the k-group partition of the data that will optimize this criterion. In theory, of course, the problem is simple; to quote Dr. Idnozo Hcahscror-Tenib, that super galactian hypermetrician who appeared in Thorndike's 1953 Presidential address to the Psychometric society, 'Is easy. Finite number of combinations Only 563 billion billion billion. Try all. Keep best'. In practice the size of n will not allow complete enumeration even using the fastest computer available since, for example, for $n = 19$, $k = 8$ there are 1,709,751,003,480 distinct partitions. This difficulty has led to the development of algorithms designed to search for a local optimum of the criterion by rearranging existing partitions and only keeping the new arrangement if it improves the criterion value. Such procedures are generally known as *hill climbing* algorithms. They begin with some arbitrary partition of the data into the required number of groups and then consider each individual one by one to see whether a move into another group produces an improvement in the current criterion value. If it does the entity is included in the other cluster and the procedure repeated until no move of a single individual causes any further improvement. The whole procedure is sometimes repeated from a different initial partition in the hope that an improved solution will be obtained. With well structured data different starting values will usually lead to the same

final set of clusters, although in general there is no way of knowing if the particular criterion value obtained is the global or merely a local optimum.

Other optimization algorithms which have been suggested in respect of the trace (W) criterion are discussed by Jensen (1968) and Gordon and Henderson (1977). The former author describes a dynamic programming algorithm which although giving a mathematically attractive statement of the problem does not seem to offer realistic practical solutions. Gordon and Henderson derive an algorithm based on the classical technique of steepest descent which in some cases performs very poorly but in others gives results which compare favourably with those obtained by other algorithms. A 'hybrid' algorithm also described by these authors does however appear to the worthy of further consideration for minimizing trace (W) and may be capable of extension to the optimization of other criteria.

Scott and Symon (1971), in an investigation of the det (W) criterion for clustering, found that problems could arise with the hill-climbing algorithm when the actual structure of the data consisted of groups of rather disparate sizes. In such cases it was found that this criterion had a tendency to provide solutions having approximately equal sized groups, with the result that the smaller group failed to be correctly identified.

3.4. Choosing the Number of Groups

Choice of criterion and choice of optimization algorithm do not exhaust the problems of this type of clustering technique. We still need to consider the formidable problem of choosing an appropriate value of k, the number of groups. The importance and difficulty of this problem have been noted by many authors including Ling (1971) and Sneath and Sokal (1973) and an early attempt at its solution was made by Thorndike (1953) who plotted average within-cluster distance against number of groups; with every increase in k there will, of course, be a decrease in this measure, but Thorndike suggests that a sudden marked flattening of the curve at any point indicates a distinctively 'correct' value for k, since, intuitively, such a point will occur when the number of groups uniquely corresponds to the configuration of points and there is relatively little gain from further increase in k. Thorndike makes some attempt to test this procedure empirically using artificial data generated to contain four clusters. Unfortunately the derived curves provide little support for this intuitive notion. Despite this a similar procedure has been advocated by other authors—the classification criterion is plotted against the number of groups and, according to Gower (1975), 'a sharp step in this plot indicates the number of classes; otherwise there is no justification for having more than one class'. In practice however the decision over whether such plots contain the necessary 'sharp step' is likely to be exceedingly subjective and in many applications of clustering this author has not found such a procedure particularly helpful.

A less subjective but still essentially informal approach to the problem is taken by Marriot (1971). In an interesting and informative discussion of the det (W) clustering criterion, he suggests that a possible criterion for assessing number of groups is to take that value of k for which $k^2|W|$ is a minimum. For unimodal distributions the minimum value is likely to give $k = 1$, for strongly grouped distributions the minimum will indicate the appropriate value of k, while for a uniform distribution the criterion should remain constant. Some simulation results given in the paper are likely to be very useful to investigators attempting to decide on a value for k, and although Marriot's results in no way provide an exact significance test for the presence of clusters they are generally very helpful in practical situations.

Some authors have attempted to derive more formal tests of number of clusters. For example, Beale (1969) gives an 'F test' which he suggests may be used to test whether a sub-

division into k_2 clusters is significantly better than a sub-division into some smaller number of clusters k_1. Experience with this statistic suggests that it will only be useful when the clusters are fairly well separated and hyper-spherical. Englemann and Hartigan (1969) give percentage points of a test for clusters based on the ratio of between groups to within groups sum of squares. In association with the multivariate mixture approach to clustering, Wolfe (1971) derives a likelihood ratio test for assessing the hypothesis that the data arises from a k_1 component mixture against the alternative that they arise from a mixture with k_2 components. Binder (1978) has criticised this test on the grounds that the proposed likelihood ratio test criterion does not necessarily have the assumed asymptotic chi-squared distribution.

Again in a discussion of the mixture approach as a model for clustering, Day (1969) suggests a test that the data is drawn from a single multivariate normal distribution against that of a mixture of two multivariate normal distributions with the same variance-covariance matrix, may be based on the maximum likelihood estimate of the generalized distance

$$\Delta = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \sum^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) .$$

Following this suggestion and using Monte Carlo techniques Everitt (1978b) has studied the null distribution of this estimate and derived significance points for such a test. Such tests are likely to be *most* useful where multivariate normality is a reasonable assumption. Some further possible approaches to the number of clusters problem are discussed in Lennington and Flake (1975).

Overall the problem of determining the most appropriate number of clusters for a set of data can be a difficult one. Despite the numerous attacks on the problem in the literature it must be said that no completely satisfactory solution is available. The main difficulties with deriving formal significance tests in this area appear to be the specification of a suitable null hypothesis, the determination of the sampling distribution of the distance or similarity measures used and the development of a flexible test procedure. Perhaps the problem is, in fact, *incapable* of any formal solution in a truly general sense simply because there is no universally acceptable definition of the term cluster. Of course, it might be argued that for practical purposes such formal significance tests are unnecessary since the investigator would do better to consider the possibility of several alternative classifications, each reflecting a different aspect of the data. Gnanadesikan and Wilk (1969) seem to be making just this point in a slightly different context when they argue that interpretability and simplicity are important in data analysis and any rigid inference of optimal number of groups (dimensions in original discussion) in the light of the observed value of a numerical index of goodness of fit, may not be productive.

3.5. Dissection and Classification

Perhaps the most difficult problem facing the user of cluster analysis techniques in practice is the assessment of the stability and validity of the clusters found by the numerical technique used. A number of questions need to be asked *and* satisfactorily answered before any given typology can be offered as a reasonable and useful system of classification. Amongst such questions are 'Do the same types emerge when new variables are used?', 'Do the same types emerge when a new sample of similar individuals is used?', 'Do the members of different groups differ on variables other than those used in deriving them?' and, in certain situations, more specific questions such as 'Do the members of the different groups respond differently to the same treatment?' However, in many reported clustering applications little consideration appears to be given to such questions; many users simply report the results of

applying one particular cluster method to a set of data and little else. Although such an approach may be thought to be suitable where the investigator is simply interested in using clustering to provide some summary and description of his data, it is obviously inadequate if he wishes to propose that the groups found are of particular importance in his area of study.

The difficulty here essentially concerns whether the user is interested in dissection or classification. Many authors, for example, Fleiss *et al.* (1971) assume that groupings determined by arbitrarily splitting a homogeneous sample, i.e. dissection, are not what are required of cluster analysis. In such a situation the ideal answer would be a technique which actually indicated that the data did not contain clusters. (This is, of course, related to the previous discussion concerning the number of groups problem). Cormack (1971) takes a similar view to Fleiss *et al.*, suggesting that classification is a technique for generating hypotheses whereas dissection is not, and where there are no distinct clusters the data will have been forced into a straight jacket which restricts the domain of possible hypotheses and makes it likely that some will be generated by the fact of dissection rather than by the data. However, other authors, for example, Ross (1971), argue that dissection is a useful activity both in everyday life and in scientific research, and the purpose of clustering should be to provide a sound basis for dissection, making use of any natural breaks that occur. Such a viewpoint is probably only reasonable when the investigator cares not at all about the relative isolation of clusters, but only about their internal homogeneity. An area where this might be appropriate is where cluster analysis techniques are used for stratification in sample surveys (see, for example, Golder and Yeomans 1973, Dahmström and Hagnell 1974 and Holgersson 1975). Perhaps the important point is that many users of clustering are not clear whether they are interested in dissection or classification, and are not helped (and in some cases apparently not bothered) by the lack of a satisfactory test for distinct clusters.

4. Cluster Analysis—Some Possible Future Developments

In this section I would like to speculate briefly on those areas which might offer potential in the future development of cluster analysis. Not having the advantages of a Madame Sosostris, such speculations may of course turn out to be wide of the mark. Nevertheless it is hoped that they will prove of use to at least some readers.

4.1. Number of Clusters Problems

This problem has been discussed in some detail in the previous section. For reasons stated there, it is probably not capable of any definitive solution; nevertheless it still presents interesting possibilities for research. Although attempts have been made to deal with the problem analytically (see, for example, Ling 1972), the most fruitful approach may be by the use of Monte Carlo techniques, as in the papers of Englemann and Hartigan (1969), Marriot (1971), and Everitt (1978*b*). In this way the null distributions of many clustering criterion could be studied and used to provide guidelines for the existence or otherwise of distinct clusters. In the same way rules such as those proposed by Mojena (1977) for deciding on the 'best' number of clusters could be empirically evaluated.

4.2. Choosing a 'Best' Clustering Method

The increasing number of cluster analysis methods available has led several authors to consider the perplexing problem of choosing a 'best' method in some sense. Fisher and Van Ness (1971), for example, whilst not considering this problem to be defined well enough

for a complete solution, suggest various admissibility conditions which they suggest will eliminate obviously bad clustering algorithms. The work of Jardine and Sibson referred to in the previous section also leads to recommendations regarding which techniques are acceptable and which are not. Whilst such theoretical approaches to this problem may be illuminating in many respects, they have not led to results acceptable in practice, and it appears unlikely that the relations between different methods and data types will be untangled solely by formal analysis and argument. An alternative and very promising approach to understanding and evaluating the variety of clustering techniques available is to compare the effectiveness of different methods across a variety of data sets. Several studies of this type have already been undertaken and referred to in the previous section. However, there is a need for more such investigations using other techniques and a greater range of data types.

4.3. Applying the 'Jackknife' to Clustering

Over the last few years the *jackknife* has been and still remains a topic of great interest in the statistical literature. The method depends on the qualitative idea that some aspects of the stability of an estimate can be judged empirically by examining how much the estimate changes as observations are removed. It has been used in a variety of contexts including the estimation of error rates in discriminant function analysis (see Lachenbruch 1975), and it might perhaps be possible to use it to evaluate clustering techniques, and methods for assessing number of clusters.

4.4. Interaction Between Cluster Methods and Informal Graphical Techniques

The last five years has seen the development of many new graphical techniques capable of being applied to multivariate data—a description of the majority of these is available in Gnanadesikan (1977). Most research workers with complex multivariate data to analyse have, as yet, little experience with the more recent of these methods, but hopefully during the near future this situation will change, and such graphical techniques will be welcomed as useful additions to the tools of the data analyst. In particular they may be very helpful when used in association with clustering methods, as an aid in the interpretation and presentation of results. Such interaction between clustering and graphical techniques may be made even more attractive by the development and increasing availability of the type of interactive computer systems described by Ball and Hall (1970) and by Tukey, Fisherkeller and Friedman (1975).

4.5. Computer Packages for Cluster Analysis

The most comprehensive computer package for cluster analysis is, undoubtedly, CLUSTAN, developed during the late 1960's and early 1970's by Dr. David Wishart. This package includes a large number of clustering techniques and a variety of distance and similarity measures. It has gained wide acceptance and is currently used by research workers in many fields. Nevertheless there is still probably room for other clustering packages providing alternative and/or additional features. One possibility would be a package based on the programs listed in Anderberg (1973) or Hartigan (1975). A further possibility would be a cluster package including as options a number of the graphical techniques described in Gnanadesikan (1977).

4.6. Improved Algorithms for Particular Techniques

Recently a number of authors have published accounts of improved algorithms for a number of cluster methods. For example, Sibson (1973) described an optimally efficient algorithm for single linkage, making its application feasible for a number of individuals well into the range 10^3 to 10^4. Defay (1977) has given a similar algorithm for the complete link method. This technique of cluster analysis has also been considered by Hansen and Delattre (1978) who show that it is reducible to the problem of optimally coloring a sequence of graphs, and derive an efficient algorithm for its implementation using this concept. The algorithm suggested by Gordon and Henderson (1977) for minimization of trace (W) has already been referred to in the previous section, and consideration should be given to a similar approach to the optimization of other clustering criterion, as an alternative to the more usual hill-climbing type of algorithm.

The development of more efficient algorithms for a variety of clustering techniques would appear to be an area of some potential. However, if such improved algorithms are to be of general use they must be incorporated rapidly into established cluster packages.

Other developments will, of course, occur with the use of cluster analysis in new areas, although with its use already noted by Wishart (1978) for classifying puberty rites of American Indian Tribes, and extracts from Plato and Jane Austin, one wonders in what other fields it can possibly emerge! No doubt theoretical developments will also take place with the increased interest of mathematicians in the area. For example, the recent publication of a Bayesian approach (see Binder 1978) has raised interesting points concerning the technique based on minimization of det (W) originally proposed as a cluster method some ten years ago. Perhaps one might also hope that a further development will be a more critical approach by users, with more than merely 'lip-service' being paid to the evaluation of solutions.

5. Summary

The 1960's saw a massive increase in the literature of cluster analysis, and a tendency for research workers in many fields to be carried along on a growing tide of euphoria for the techniques. Fortunately the 1970's has seen this tendency less in evidence, partly because of the appearance of papers openly critical of the attitude prevalent earlier of seemingly regarding clustering as an easy alternative to being forced to sit and think. (The major example of such a paper is the review by Cormack 1971). Most (although by no means all) investigators are now more wary of the whole area, having become aware of the varied and difficult problems facing the cluster analysis user in practice. This more critical approach is to be welcomed and should lead to more worthwhile results being produced in the future than have often been produced in the past.

Résumé

Le nombre des techniques de classification a dramatiquement augmenté depuis 10 ou 15 ans, et elles ont été utilisées dans des domaines aussi différents les uns des autres que l'archéologie et al psychiatrie. Cependant, beaucoup de problèmes ne sont pas résolus et les utilisateurs potentiels de ces méthodes ont besoin d'en avoir connaissance, s'ils ne veulent pas faire face à des résultats inintéressants ou même trompeurs. On débat ici de ces problèmes et on tente d'indiquer des possibilités de travaux futurs dans ce domaine.

References

Anderberg, M. R. (1973). *Cluster Analysis for Applications* Academic Press, New York.

Ball, G. H. and Hall, D. J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics 12*, 17–31.

Beale, E. M. L. (1969). Euclidean cluster analysis. *Bulletin of the International Statistical Institute 43*, 92–94.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika 65*, 31–38.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin 83*, 377–388.

Cattell, R. B. (1952). The three basic factor-analytic research designs—their inter-relations and derivatives. *Psychological Bulletin 49*, 499–520.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A 134*, 321–367.

Cox, D. R. (1978). Some remarks on the role in statistics of graphical methods. *Applied Statistics 27*, 4–9.

Cunningham, K. M. and Ogilvie, J. C. (1972). Evaluation of hierarchical grouping techniques: A preliminary study. *The Computer Journal 15*, 209–213.

Dahmström, P. and Hagnell, M. (1974). The formation of strata using cluster analysis. *Research Report No. 4*, Department of Statistics, Lund, Sweden.

D'Andrade, R. G. (1978). U-statistic hierarchical clustering. *Psychometrika 43*, 59–67.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*, 463–474.

Defay, D. (1977). An efficient algorithm for a complete link method. *Computer Journal 20*, 364–366.

Englemann, L. and Hartigan, J. A. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association 64*, 1647–1648.

Everitt, B. S. (1974). *Cluster Analysis.* Heinemann, London.

Everitt, B. S. (1976). Cluster analysis. In *The Analysis of Survey Data,* Vol. 1. C. A. O'Muircheertaigh and C. Payne (eds.) Wiley and Son, New York.

Everitt, B. S. (1978*a*). *Graphical Techniques for Multivariate Data.* Heinemann, London.

Everitt, B. S. (1978*b*). A test of multivariate normality against the alternative that the distribution is a mixture. Submitted to *Biometrics.*

Fisher, L. and Van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika 58*, 91–104.

Fleiss, J. L., Lawlor, W., Platman, S. R. and Fieve, R. R. (1971). On the use of inverted factor analysis for generating typologies. *Journal of Abnormal Psychology 77*, 127–132.

Fleiss, J. L. and Zubin, J. (1969). On the methods and theory of clustering. *Multivariate Behavioural Research 4*, 235–250.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association 62*, 1159–1178.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

Gnanadesikan, R. and Wilk, M. B. (1969). Data analytic methods in multivariate statistical analysis. In *Multivariate Analysis* II. P. R. Krishnaiah (ed.). Academic Press, New York.

Golder, P. A. and Yeomans, K. A. (1973). The use of cluster analysis for stratification. *Applied Statistics 22*, 213–219.

Gordon, A. D. and Henderson, J. J. (1977). An algorithm for Euclidean sum of squares classification. *Biometrics 33*, 355–362.

Gower, J. C. (1975). Goodness-of-fit criteria for classification and other patterned structures. In *Procedure of the 8th International Conference on Numerical Taxonomy*, W. H. Freeman and Co., 38–62.

Hansen, P. and Delattre, M. (1978). Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association 73*, 397–403.

Hartigan, J. A. (1975). *Clustering Algorithms.* Wiley and Son, New York.

Holgerson, M. (1975). Multivariate Stratification with the use of cluster analysis. *Research Report, Department of Statistics,* University of Uppsala, Sweden.

Hubert, L. J. (1973). Monotone invariant clustering procedures. *Psychometrika 38*, 47–62.

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal 11*, 117–184.

Jensen, R. E. (1968). A dynamic programming algorithm for cluster analysis. *Operational Research 17,* 1034–1056.
Johnson, S. C. (1967). Hierarchial clustering schemes. *Psychometrika 32*, 241–254.
Kuiper, F. K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics 31*, 777–783.
Lachenbruch, P. A. (1975). *Discriminant Analysis.* Hafner Press, New York.
Lennington, R. K. and Flake, R. H. (1975). Statistical evaluation of a family of clustering methods. In *Procedure of the 8th International Conference on Numerical Taxonomy*, W. H. Freeman and Co.
Ling, R. F. (1971). Cluster analysis. Unpublished Ph.D. thesis, Department of Statistics, Yale University, New Haven, Connecticut.
Ling, R. F. (1972). On the theory and construction of *k*-clusters. *Computer Journal 15*, 326–332.
Maronna, R. and Jacovkis, P. M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics 30*, 499–505.
Marriot, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics 27*, 501–514.
McRae, D. J. (1971). MICKA, a Fortran IV iterative K-means cluster analysis program. *Behavioural Science 16*, 423–424.
Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal 20*, 359–363.
Paykel, E. S. and Rassaby, E. (1978). Classification of suicide attempters by cluster analysis. *British Journal of Psychiatry 133*, 45–52.
Ross, G. (1971). Discussion of Cormack—A review of classification. *Journal of the Royal Statistical Society, Series A 134*, 321–367.
Scott, A. J. and M. Knott (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics 30*, 507–512.
Scott, A. J. and Symon, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics 27*, 387–398.
Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method. *Computer Journal 16*, 30–34.
Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy.* W. H. Freeman and Co., San Francisco.
Thorndike, R. L. (1953). "Who belongs in a family?" *Psychometrika 18*, 267–276.
Tukey, J. W., Fisherkeller, M. A. and Friedman, J. H. (1975). PRIM-9: An interactive multidimensional data display and analysis system. *Proceedings of the 4th International Congress for Stereology, Sept. 4-9.* Gaithersburg, Maryland.
Ward, J. H. (1963). Hierarchial grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 236–244.
Williams, W. T., Lance, G. N., Dale, M. B. and Clifford, H. T. (1971). Controversy concerning the criteria for taxonometric strategies. *Computer Journal 14*, 162–165.
Wishart, D. (1969). Mode analysis. In *Numerical Taxonomy.* A. J. Cole (ed.), Academic Press, New York, 282–308.
Wishart, D. (1978). *Clustan User Manual*, 3rd ed. Program Library Unit, Edinburgh University, Edinburgh, Scotland.
Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research 5*, 329–350.
Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distribution. In *Naval Personnel and Training Research Laboratory Technical Bulletin* STB, San Diego, California, 72–2.
Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing Gestalt clusters. *IEEE Transactions on Computers C20*, 68–86.
Zubin, J. and Fleiss, J. L. (1965). Taxonomy in the mental disorders—a historical perspective. In *Symposium on Explorations in Typology with Special Reference to Psychotics.* Human Ecology Fund, New York.

References

Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics
Geoffrey H. Ball; David J. Hall
Technometrics, Vol. 12, No. 1. (Feb., 1970), pp. 17-31.
Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28197002%2912%3A1%3C17%3ASIOIGC%3E2.0.CO%3B2-4

Bayesian Cluster Analysis
D. A. Binder
Biometrika, Vol. 65, No. 1. (Apr., 1978), pp. 31-38.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197804%2965%3A1%3C31%3ABCA%3E2.0.CO%3B2-Q

Some Remarks on the Role in Statistics of Graphical Methods
D. R. Cox
Applied Statistics, Vol. 27, No. 1. (1978), pp. 4-9.
Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281978%2927%3A1%3C4%3ASROTRI%3E2.0.CO%3B2-T

Estimating the Components of a Mixture of Normal Distributions
N. E. Day
Biometrika, Vol. 56, No. 3. (Dec., 1969), pp. 463-474.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28196912%2956%3A3%3C463%3AETCOAM%3E2.0.CO%3B2-B

LINKED CITATIONS

A Method for Plotting the Optimum Positions of an Array of Cortical Electrical Phosphenes
B. S. Everitt; D. N. Rushton
Biometrics, Vol. 34, No. 3. (Sep., 1978), pp. 399-410.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197809%2934%3A3%3C399%3AAMFPTO%3E2.0.CO%3B2-Z

Admissible Clustering Procedures
Lloyd Fisher; John W. Van Ness
Biometrika, Vol. 58, No. 1. (Apr., 1971), pp. 91-104.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197104%2958%3A1%3C91%3AACP%3E2.0.CO%3B2-%23

On Some Invariant Criteria for Grouping Data
H. P. Friedman; J. Rubin
Journal of the American Statistical Association, Vol. 62, No. 320. (Dec., 1967), pp. 1159-1178.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196712%2962%3A320%3C1159%3AOSICFG%3E2.0.CO%3B2-D

The Use of Cluster Analysis for Stratification
P. A. Golder; K. A. Yeomans
Applied Statistics, Vol. 22, No. 2. (1973), pp. 213-219.
Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281973%2922%3A2%3C213%3ATUOCAF%3E2.0.CO%3B2-0

An Algorithm for Euclidean Sum of Squares Classification
A. D. Gordon; J. T. Henderson
Biometrics, Vol. 33, No. 2. (Jun., 1977), pp. 355-362.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197706%2933%3A2%3C355%3AAAFESO%3E2.0.CO%3B2-O

Complete-Link Cluster Analysis by Graph Coloring
Pierre Hansen; Michel Delattre
Journal of the American Statistical Association, Vol. 73, No. 362. (Jun., 1978), pp. 397-403.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197806%2973%3A362%3C397%3ACCABGC%3E2.0.CO%3B2-C

LINKED CITATIONS

391: A Monte Carlo Comparison of Six Clustering Procedures
F. Kent Kuiper; Lloyd Fisher
Biometrics, Vol. 31, No. 3. (Sep., 1975), pp. 777-783.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197509%2931%3A3%3C777%3A3AMCCO%3E2.0.CO%3B2-W

Multivariate Clustering Procedures with Variable Metrics
Ricardo Maronna; Pablo M. Jacovkis
Biometrics, Vol. 30, No. 3. (Sep., 1974), pp. 499-505.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197409%2930%3A3%3C499%3AMCPWVM%3E2.0.CO%3B2-4

A Cluster Analysis Method for Grouping Means in the Analysis of Variance
A. J. Scott; M. Knott
Biometrics, Vol. 30, No. 3. (Sep., 1974), pp. 507-512.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197409%2930%3A3%3C507%3AACAMFG%3E2.0.CO%3B2-C

Clustering Methods Based on Likelihood Ratio Criteria
A. J. Scott; M. J. Symons
Biometrics, Vol. 27, No. 2. (Jun., 1971), pp. 387-397.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197106%2927%3A2%3C387%3ACMBOLR%3E2.0.CO%3B2-1

Hierarchical Grouping to Optimize an Objective Function
Joe H. Ward, Jr.
Journal of the American Statistical Association, Vol. 58, No. 301. (Mar., 1963), pp. 236-244.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196303%2958%3A301%3C236%3AHGTOAO%3E2.0.CO%3B2-9