Quality and Quantity, 14 (1980) 75–100

CLUSTER ANALYSIS

BRIAN EVERITT

Institute of Psychiatry, University of London, U.K.

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. 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 those 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 more commonly *cluster analysis methods*. During the last two decades a vast variety of such techniques have been developed, and comprehensive reviews of the area are available in Cormack (1971) and Everitt (1974). A further brief review including more recent work is given in the next section, and this is followed by a discussion of some of the unresolved problems of this type of analysis. Next a numerical example is given in which some recently-developed methods are

used and finally some attempt is made to indicate possible future developments in the area.

Developments since 1960

Prior to the late 1950's the literature of cluster analysis and the number of clustering techniques were of a fairly manageable size. Zubin (1938) introduced a method for measuring the similarity of people's behaviour. Tyron (1939) proposed a number of clustering methods whose logic was related to that of factor analysis. Thorndike (1953) in an amusing Presidential address to the Psychometric Society described a type of cluster algorithm similar in many respects to the 'k-means' approach popular today. Single linkage clustering was first proposed in a paper by Florek et al. (1951), and later described by McQuitty (1957) and Sneath (1957), and Sokal and Michener (1958) introduced group average and centroid clustering.

Since 1960 however there has been what can only be termed an explosion in the area with an almost exponential growth in both the literature and the number of methods and algorithms. This dramatic increase in interest was largely a result of the increased availability of the electronic computer which was generally necessary to take the burden of the onerous amount of computation required by many cluster methods. A further reason for the growth of this type of analysis at this time was the publication in 1963 of Sokal and Sneath's pioneering work *Principles of Numerical Taxonomy*. This spelt out clearly some of the general problems of numerical classification, and became a major influence in the area, particularly of course in the biological field.

The 1960's saw the development of a number of cluster methods all essentially based on the idea of making clusters as homogeneous as possible by minimizing the sum of squared deviations of observations from their cluster means, a criterion usually referred to as the *error sum of squares*. Ward (1963) described an *agglomerative* technique implementing this idea, while Edwards and Cavalli-Sforza (1965) suggest an algorithm of the *divisive* type. (See Everitt, 1974, Chapter 2 for a description of the terms in italics.) The error sum of squares criterion was also considered by several other authors including Forgey (1964), Jancey (1966), MacQueen (1967) and Ball and Hall (1967). It does however suffer from the disadvantage that the solutions produced are

not invariant with respect to changes of scale in the variables and this led Friedman and Rubin (1967) to suggest several other methods whose solutions *are* invariant under such transformations of the raw data. The advantages of these techniques will be considered in the next section.

In the mid 1960's a series of papers, Lance and Williams (1967a, 1967b) discussed general properties of both hierarchical and non-hierarchical clustering methods. They introduced a recurrence formula which aided considerably the computer implementation of the commonly used hierarchical techniques such as single-linkage, complete-linkage, group average and centroid clustering. Wishart (1969a) extended this formula to include Ward's method based on the error sum of squares, and introduced the first comprehensive clustering software, namely CLUSTAN I (see Wishart, 1969b).

The late 1960's saw a number of interesting theoretical developments. For example, papers by Hartigan (1967) and Johnson (1967) gave similar characterizations of a hierarchic dendrogram, and Johnson showed that the requirement that clustering should be invariant with respect to monotone transformations of the similarity or distance measures leads directly to the single and complete-linkage methods. In a similar vein Jardine and Sibson (1968) attempted to develop a rigorous mathematical theory for clustering and the framework they constructed led them to reject all but the single-linkage method among the well-known hierarchical techniques and to the development of a mathematically acceptable method allowing overlapping clusters. A detailed account of the approach taken by Jardine and Sibson is given in their text *Mathematical Taxonomy* (1971), and some further discussion of the issues raised is given in the next section.

An attempt to formulate the cluster problem in terms of an acceptable statistical model was made in a series of papers by Wolfe (1965, 1967, 1970), and a similar approach was considered by Day (1969). Both these authors suggested using a "mixture" approach to cluster analysis. Mixture distributions have been the subject of much investigation over the years; see, for example, Pearson (1894), Charlier and Wicksell (1924) and Cohen (1967), and a comprehensive review by Holgerson and Jorner (1976). Day and Wolfe consider the maximum likelihood estimation of the parameters in a mixture of multivariate normal distributions, and Wolfe has produced a program, NORMIX, for solving the maximum likelihood equations, which has enabled this approach to be used in a number of cluster applications including Everitt et al. (1971) and Wolfe (1978).

The early 1970's saw the appearance of a number of reviews of cluster analysis. Notable amongst these were the papers by Ball (1971) and Cormack (1971). Cormack was frankly critical of the whole area, and especially of the growing tendency by some researchers to regard clustering as a panacea and an easy alternative to being forced to sit and think. A little later several texts wholly denoted to cluster analysis were produced, for example, Anderberg (1973), Everitt (1974) and Hartigan (1975), marking perhaps a consolidation phase in the cluster analysis literature.

Recent developments in the field of cluster analysis include a method described by Hartigan (1972) for the *simultaneous* clustering of individuals *and* variables, an interesting and informative account of the metrics used in cluster analysis by Maronna and Jacovkis (1974), a method for deciding on the best number of groups proposed by Mojena (1977), a Bayesian approach to clustering suggested by Binder (1978), and a reformulation of the complete-linkage clustering method as a problem of optimally colouring a sequence of graphs by Hansen and Delattre (1978). A number of these, and other recent work in the area, will be discussed further in the next section. A further recent development relevant to cluster analysis users is the upsurge of interest in graphical techniques, particularly those applicable to multivariate data. A number of these, for example, Andrews (1972) and Chernoff (1973) have been used in clustering problems and a full account of the area is available in Gnanadesikan (1977) and Everitt (1978a).

Some Unresolved Problems

There are a large number of problems associated with clustering techniques. For example, 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 be used? Here we shall discuss just a few of these problems and some of the more recent attempts to overcome them.

Hierarchical clustering

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 produced by a hierarchical clustering method resemble 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 these biological applications questions concerning the optimal number of groups do not arise, for 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 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 characterized by an ultrametric. Since the input similarities or distances are not generally ultrametric (and only occasionally metric), Jardine and Sibson (1968) argue 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 are reasonable for any such transformation to satisfy, and prove that single-linkage is the only method satisfying all their criteria. The implication apparently is 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. They 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. A number of such empirical studies have been undertaken for example those reported by Cunningham and Ogilvie (1972), Kuiper and Fisher (1975) and Blashfield (1976). Although no consistent conclusions were reached of the type that would enable us to declare that cluster method A was superior in all situations, it was noticeable that the mathematically respectable single-linkage technique was for the most part, the *least* successful method for the data sets used. Such empirical studies can, of course, never afford a complete evaluation of cluster 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 this method does have a number of desirable properties, perhaps the most important of which is that its results are invariant under monotone 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 retaining its useful mathematical properties but making it more practically relevant. Examples are the methods proposed by Wishart (1969) and Zahn (1971). In addition Sibson (1973) has produced a very efficient algorithm for the single-link technique which enables it to handle very large data sets and this may be regarded as a distinct advantage in many practical situations. (Defay, 1977, has given a similar algorithm for the complete link method.)

Partitioning into groups

Let us now move on to consider those clustering techniques which seek to partition the data into k groups so as to optimize some predefined numerical measure. The measure is defined so that extreme values (high for some measures, low for others), are considered indicative of a desirable set of clusters. Such methods differ from those discussed in the previous paragraph in that the solution does not portray hierarchical relationships among the entities. The clusters obtained from such methods 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 of k will be discussed in detail later.

Numerical criteria

Several numerical criteria have been proposed for this approach to clustering. One already mentioned in the previous section is the pooled within-cluster error sum of squares, i.e., trace (W) where W is the (p × p) matrix of within-cluster sums of squares and products. Solutions with low values of this criterion are indicative of a homogeneous set of clusters. Algorithms to minimize trace (W) are described 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 when applying the technique to the raw data

and, say, the data standardized in the usual way (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 unavoidable. A further problem with the min {trace (W)} criterion is that the clusters produced are constrained to being hyperspherical. In cases where the real clusters in the data are of some other shape this may produce misleading solutions. Examples are given in Wishart (1969b) 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, which is 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

$$\prod_{i=1}^{k} |W_i|^{n_i} \qquad (1)$$

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

$$\min \left(p \sum_{i} (n_i - 1) |W_i|^{1/p} \right) \qquad (2)$$

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

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 503 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 all individuals one by one to see whether moving an individual 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 be 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.

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 solutions.

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 hyperspherical. 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 criticized 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 that 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

$$\Delta = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \qquad (3)$$

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 pos-

sible 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 that any rigid inference of optimal number of groups (dimensions in original discussion) from the observed value of a numerical index of goodness-of-fit may not be productive.

Stability of solution

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 is 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 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. Where there are no distinct clusters the data will have been forced into a strait-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 that 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. One area where this might be appropriate is where cluster analysis techniques are used for stratification in sample surveys (for example, see Golder and Yeomans (1973), Dahmström and Hagnell (1974) and Holgerson (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 by the lack of satisfactory tests for distinct clusters.

Numerical Example

As an example of the use of cluster analysis in practice we shall employ a set of data first considered by Powell, Clark and Bailey (1978). My thanks are due to Dr Graham Powell for allowing me to use this data set. These data consisted of 86 cardio-vascular accident cases referred to

hospital for the assessment of speech functions. All subjects were administered the Minnesota Test for the differential diagnosis of Aphasia, which consists of a series of 43 tests designed to chart the deficits and residual capacities of language-impaired subjects. Details may be found in Schuell (1965) and Schuell and Sefer (1973).

The first analysis performed by Powell et al. on these data used not the original 43 test scores but four derived section scores measuring (1) auditory disturbance, (2) visual and reading disturbance, (3) speech and language disturbance and (4) visuo-motor and writing disturbance. By applying the minimization of det (W) clustering technique mentioned in the previous section they found four groups of subjects which they labelled as Severe, High-Moderate, Low-Moderate and Mild, and concluded that there was no evidence for a typology of aphasia other than in terms of sheer severity of symptomatology.

For the purposes of this chapter the data were reanalyzed using Day's algorithm for estimating the parameters in a mixture of multivariate normal distributions with a common variance-covariance matrix. After several runs using different starting points the best solution found (i.e., the one with greatest likelihood) had a value of $\hat{\Delta}$, the maximum likelihood estimator of the generalized distance statistic, mentioned in the previous section, of 3.349. Using the table given in Everitt (1978b) (part of which is reproduced here as Table 1), we find this to be just around the 5 per cent significance level; consequently there is some (perhaps rather weak) evidence that these data consist of two groups. The maximum likelihood estimates of the group means and the (assumed) common variance-covariance matrix are shown in Table 2. The difference between the two groups is seen to be simply that one gets high scores on all four variables, while the other gets much lower scores. This might be an even clearer demonstration than that of Powell et al. that differences between aphasic patients are differences of *degree* of symptomatology rather than differences of *pattern*. However it is also possible that this solution identifies a group of subjects whose speech functions have been affected irreversibly by their heart attack and a group who are more mildly affected and who might benefit from treatment.

The analyses just described were based upon four section scores derived from the original 43 items of Schuell's test and it is possible that these do not adequately reflect the

TABLE 1

5% Significance Points for a test of $\Delta = 0$ (reproduced from Everitt, 1978)

	k	
n	4	7
60	3.6	4.3
80	3.3	3.9
100	3.1	3.6

Points for values of n other than those given may be found by linear interpolation.

TABLE 2

Maximum likelihood estimates of means and (assumed) common variance-covariance matrix for section score data

		V1	V2	V3	V4
Group 1:	n = 38	51.4	71.9	106.9	63.4
Group 2:	n = 48	18.3	19.5	30.1	26.7

$$\hat{\Sigma} = \begin{bmatrix} 272.8 & 174.5 & 245.8 & 227.3 \\ 174.5 & 400.4 & 190.1 & 248.3 \\ 245.9 & 190.1 & 720.0 & 182.0 \\ 227.3 & 248.3 & 182.0 & 378.4 \end{bmatrix}$$

meaning of those individual items. Consequently Powell et al. carried out a factor analysis of these items and this yielded seven factors with eigenvalues greater than unity. The factor structure was virtually identical to that reported by Schuell, Jenkins and Londis (1971), the first factor accounting for some 45.9 per cent of the total variance and the next six factors accounting for between 8.8 per cent and 2.9 per cent. Each subject was now scored on each of the seven factors, and these scores used for further analyses. Powell et al. found that the cluster analysis results using the factor scores were very similar to those found using the four section scores. Application of Day's "unmixing" algorithm gave a value of $\hat{\Delta}$ of 4.278 which exceeds the 5 per cent point for $n = 86$ and $p = 7$ (found by linear interpolation in Table 1). The two groups corresponded almost exactly to those found using the four section scores, with disagreement only being found on three subjects. The estimated mean vectors and variance-covariance matrix appear in Table 3.

TABLE 3

Maximum likelihood estimates of means and (assumed) common variance-covariance matrix for factor score data

		V1	V2	V3	V4	V5	V6	V7
Group 1:	n = 38	70.9	-9.8	64.4	-60.4	-22.0	58.8	-41.3
Group 2:	n = 48	-53.7	7.8	-50.3	44.5	17.8	-60.8	30.9

$$\hat{\Sigma} = \begin{bmatrix} 5979.2 & 642.1 & -267.2 & -1330.7 & 450.3 & 812.1 & -522.0 \\ 642.1 & 9792.7 & 431.2 & -853.2 & -465.7 & 909.8 & -2649.3 \\ -267.2 & 431.2 & 6618.3 & 215.5 & -611.7 & -763.3 & 444.0 \\ -1330.7 & -853.2 & 215.5 & 7101.8 & -127.2 & 669.7 & 747.2 \\ 450.3 & -465.7 & -611.7 & -127.2 & 9558.2 & -13.8 & 726.2 \\ 812.1 & 909.8 & -763.3 & 669.7 & -13.8 & 1893.3 & -256.3 \\ -522.0 & -2649.3 & 444.0 & 747.2 & 726.2 & -256.3 & 8436.1 \end{bmatrix}$$

To obtain some graphical representation of the results Andrews plots (see Andrews, 1972), were found for the two groups of individuals for whom Day's method on the section scores and on the factor scores agreed. These plots are shown in Figures 1-4. The plots for the individuals in the two groups are not very different from each other, either for the four-variable or seven-variable analysis, and indicate that the two-group solution discussed above should perhaps be treated with a certain amount of caution. Of course, in any detailed study of these data Andrews' plots of smaller subsets of individuals would need to be studied to build up a complete picture of the structure.

Overall the results show that the data described by Powell et al. probably consists of two groups of subjects, certainly no more, and that differences are essentially of degree of symptomatology. There is certainly no evidence for the seven categories of aphasia originally proposed by Schuell and based on visual inspection of many test profiles and clinical experience.

Future Developments

The increase of interest in cluster analysis methods over the last 10 to 15 years has been dramatic. The methods have now been used in the analysis of problems as diverse as classifying puberty rites of North American Indians, studying the penile morphology of New Guinea rodents, investigating the process by which cockroaches behaviorally recover from cold stress, and assessing extracts from Plato and Jane Austen. There is now some evidence that a consolidation phase has been reached, and that more effort is being applied to the evaluation of the properties of *existing* cluster methods than to the development of new techniques. Some of this work is theoretical. For example, Fisher and Van Ness (1971) suggest various admissibility conditions against which to evaluate cluster methods, which they suggest will help to eliminate obviously bad methods. The work of Jardine and Sibson referred to above also leads to recommendations regarding which techniques are acceptable and which are not. While such theoretical approaches to the problem may be illuminating in many respects, they have not led to results of wide practical applicability and it appears unlikely that the relations between different methods and data types will be untangled solely by formal analy-

(continued on page 96)

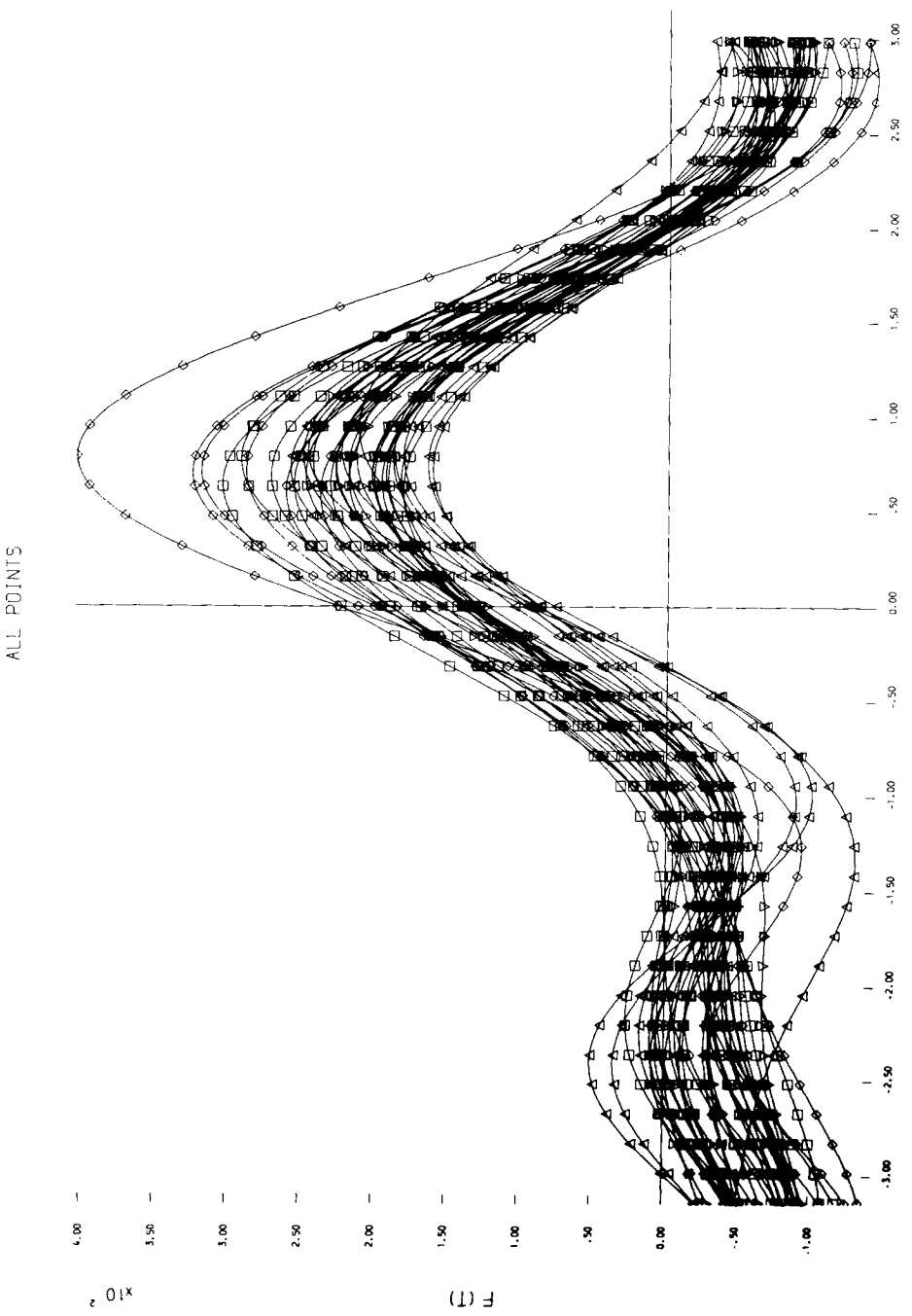

Fig. 1. Andrews plots for first cluster found from the four section scores.

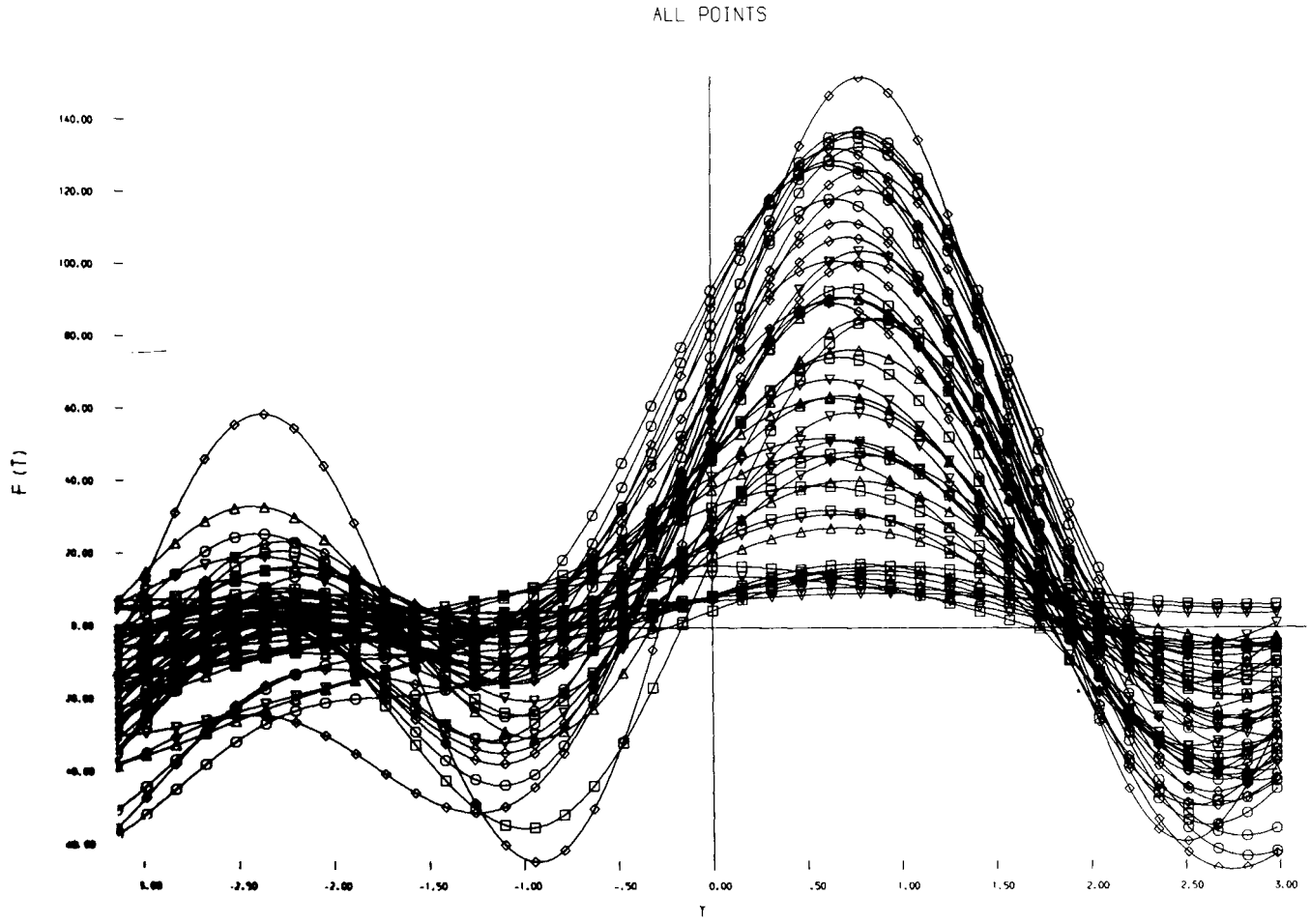

Fig. 2. Andrews plots for second cluster found from the four section scores.

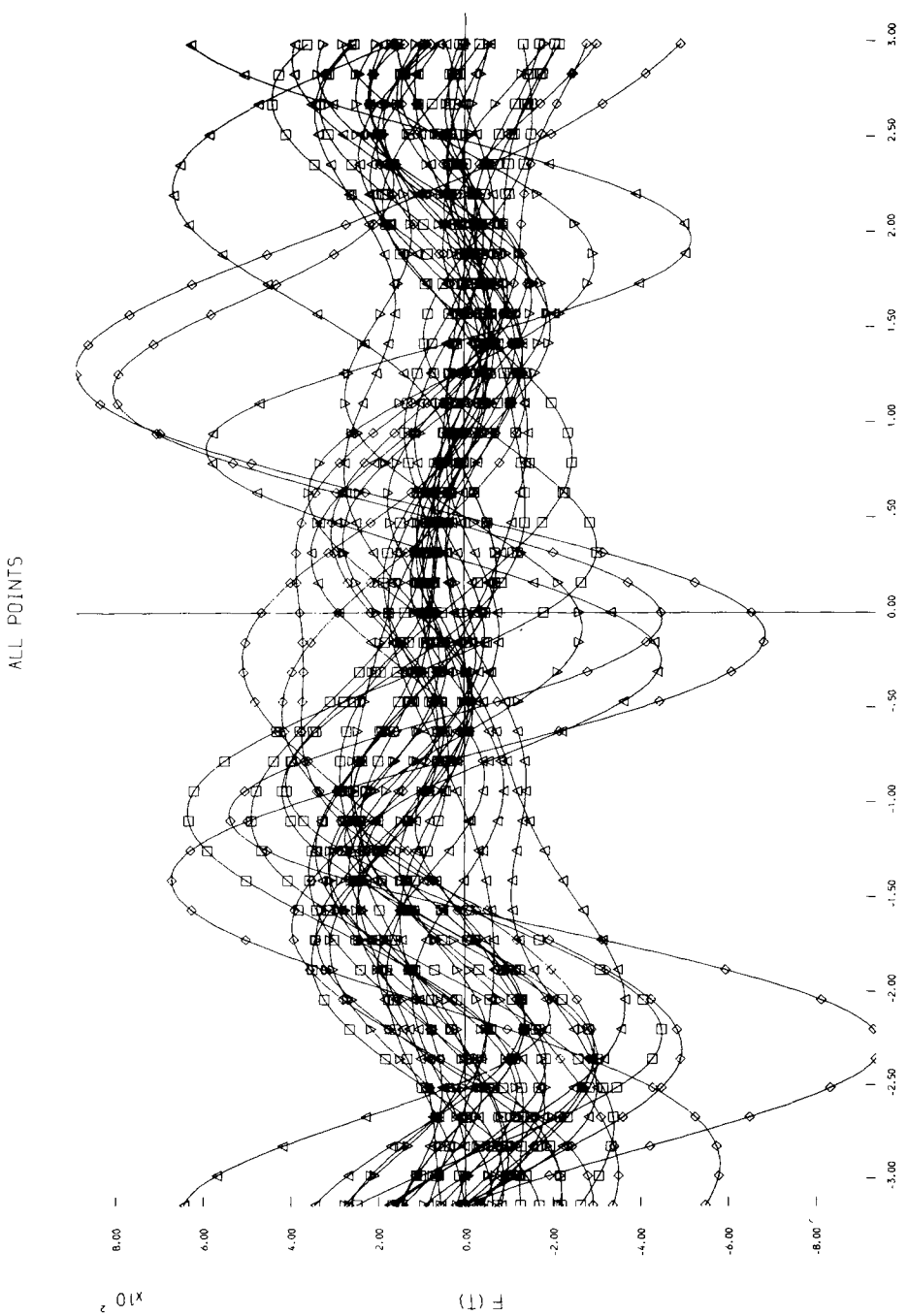

Fig. 3. Andrews plots for first cluster found from the seven factor scores.

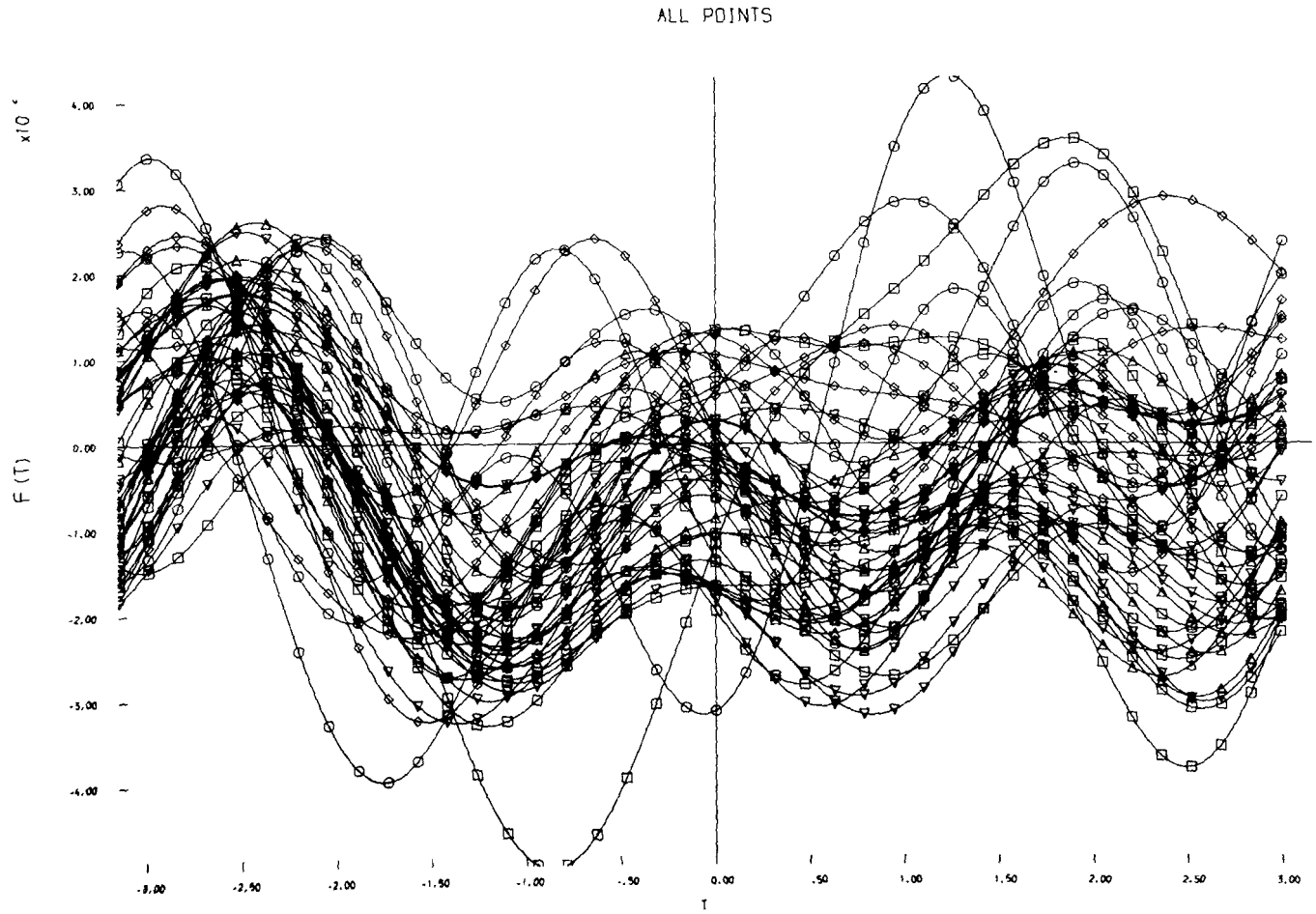

Fig. 4. Andrews plots for second cluster found from the seven factor scores.

sis and argument. Empirical studies of the type discussed above are likely to be of more practical relevance and there is a need for more of these using a greater variety both of techniques and data types.

A further area which presents interesting possibilities for research is the ever-present "number of clusters" problem. Again attempts have been made to deal with this analytically (see, for example, Ling, 1972) but it might also prove more amenable to empirical studies using Monte Carlo techniques, as in the papers of Engleman and Hartigan (1969), Marriot (1971) and Everitt (1978b).

A current growth area in statistics is that of graphical techniques, particularly those that can be applied to multivariate data. Most research workers with complex multivariate data to analyze 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. The next few years should see an increasing interaction and understanding of the relationship between clustering and graphical methods. Such interaction 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) among others.

A consumer report on *cluster analysis software* published in 1977 (see Blashfield, 1977) shows that there is considerable potential for future work in this area. The most comprehensive computer package at present available 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 similarity and distance measures. It has gained wide acceptance and is currently used by research workers in many fields. However, Blashfield's report indicates that it could be improved in several ways, and there is also probably room for other 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).

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 critical review papers such as Cormack (1971) which identified clearly the absurdities in much of the published material on clustering. 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.

References

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

Andrews, D.F. (1972). 'Plots of high dimensional data', *Biometrics* 28: 125-136.

Ball, G.H. (1971). *Classification Analysis*. Stanford Research Institute S R I Project 5533.

Ball, G.H. and Hall, D.J. (1967). 'A clustering technique for summarizing multivariate data', *Behaviour Sci.* 12: 153-155.

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', *Bull. I.S.I.* 43, Book 2: 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', *Psych. Bull.* 83: 377-388.

Blashfield, R.K. (1976). Questionnaire on Cluster Analysis Software. Unpublished Paper, Pennsylvania State University.

Blashfield, R.K. (1977). *A Consumer Report on Cluster Analysis Software*. Unpublished Research Report, Pennsylvania State University.

Charlier, C.V.L. and Wicksell, S.D. (1924). 'On the dissection of frequency functions', *Arkiv för Matematik Astronomy och Frjsik* 18, No. 6.

Chernoff, H. (1973). 'Using Faces to represent points in k-dimensional space graphically', *J. Am. Statist. Assoc.* 68: 361-368.

Cohen, A.C. (1967). 'Estimation in mixtures of two Normal Distributions', *Technometrics* 9: 15-28.

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*. Dept. 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.

Defays, D. (1977). 'An efficient algorithm for a complete link method', *Computer Journal* 20: 364-366.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965). 'A method for cluster analysis', *Biometrics* 21: 362-375.

Englemann, L. and Hartigan, J.A. (1969). 'Percentage Points of a Test for clusters', *J. Amer. Statist.*

Assoc. 64: 1647-1648.
Everitt, B.S. (1974). *Cluster Analysis*. London: Heinemann.
Everitt, B.S. (1976). 'Cluster Analysis', C.A. O'Muircheartaigh and C. Payne (eds.) *The Analysis of Survey Data*. New York: Wiley.
Everitt, B.S. (1978a). *Graphical Techniques for Multivariate Data*. London: Heinemann.
Everitt, B.S. (1978b). A Test of Multivariate Normality against the alternative that the distribution is a mixture. Submitted to *Biometrics*.
Everitt, B.S., Gourlay, A.J. and Kendall, R.E. (1971). 'An attempt at validation of traditional psychiatric syndromes by cluster analysis', *Brit. J. of Psych.* 119: 399-412.
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', *J. of Abnormal Psych.* 77: 127-132.
Fleiss, J.L. and Zubin, J. (1969). 'On the methods and theory of clustering', *Multiv. Behav. Res.* 4: 235-250.
Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H. and Zubrzycki, S. (1951). 'Sur la liaison et la division des points d'un ensemble fini', *Colloquium Math.* 2: 282-285.
Forgey, E.W. (1964). Evaluation of several methods for detecting sample mixtures from different N-dimensional populations. Amer. Psychol. Assoc., Los Angeles, California.
Friedman, H.P. and Rubin, J. (1967). 'On some Invariant criteria for grouping data', *J. Am. Statist. Assoc.* 62: 1159-1178.
Gower, J.C. (1975). 'Goodness-of-fit criteria for classification and other patterned structures', *Proceedings of the 8th International Conference on Numerical Taxonomy*: 38-62.
Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.
Gnanadesikan, R. and Wilk, M.B. (1969). 'Data analytic methods in multivariate statistical analysis', in P.R. Krishnaiah (ed.) *Multivariate Analysis II*. New York: Academic Press.
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.T. (1977). 'An algorithm for Euclidean Sum of Squares Classification *Biometrics* 33: 355-362.
Hansen, P. and Delattre, M. (1978). 'Complete-Link Cluster Analysis by Graph Coloring', *J. Amer. Statist. Assoc.* 73: 397-403.
Hartigan, J.A. (1967). 'The Representation of Similarity Matrices by Trees', *J. Amer. Statist. Assoc.* 62: 1140-1158.
Hartigan, J.A. (1972). 'Direct Clustering of a Data Matrix', *J. Amer. Statist. Assoc.* 67: 123-129.
Hartigan, J.A. (1975). *Clustering Algorithms*. New York: Wiley.
Holgerson, M. (1975). *Multivariate Stratification with the use of Cluster Analysis*. Research Report, Dept. of Statistics. Sweden: University of Uppsala.
Holgerson, M. and Jorner, U. (1976). *The decomposition of a mixture into normal components, A review*. Research Report 76-113, University of Uppsala.
Hubert, L.J. (1973). 'Monotone invariant clustering procedures', *Psychometrika* 38: 47-62.
Jancey, R.C. (1966). 'Multidimensional Group Analysis', *Aust. J. Bot.* 14: 127-130.
Jardine, N. and Sibson, R. (1968). 'The construction of hierarchic and non-hierarchic classifications', *Computer J.* 11: 117-184.
Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. New York: Wiley.
Jensen, R.E. (1968). 'A dynamic programming algorithm for cluster analysis', *Operations Res.*, 1034-1056.
Johnson, S.C. (1967). 'Hierarchical 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*. New York: Hafner.
Lance, G.N. and Williams, W.T. (1967a). 'A general theory of classificatory sorting strategies. 1 Hierarchial Systems', *Computer J.* 9: 373-380.

Lennington, R.K. and Flake, R.H. (1975). 'Statistical Evaluation of a family of clustering methods', *Proc. of the 8th International Conference on Numerical Taxonomy.*

Ling, R.F. (1971). *Cluster Analysis.* Unpublished Ph.D. Thesis, Dept. of Statistics, Yale University.

Ling, R.F. (1972). 'On the theory and construction of k-clusters', *Computer J.* 15: 326-332.

MacQueen, J. (1967). 'Some methods for classification and analysis of multivariate observations', *Proc. 5th Berkeley Symp.* 1: 281-297.

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.

McQuitty, L.L. (1957). 'Elementary linkage analysis for isolating orthogonal and oblique types and typal relevances', *Educ. Psychol. Measmt.* 17: 207-229.

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 J.* 20: 359-363.

Paykel, E.S. and Rassaby, E. (1978). "Classification of suicide attempters by cluster analysis', *Brit. J. of Psychiatry* 133: 45-52.

Pearson, K. (1894). 'Contributions to the mathematical theory of evolution', *Phil. Trans. Royal Soc.* 185: 71-110.

Powell, G.E., Clark, E. and Bailey, S. (1978). Categories of Aphasia: a cluster analysis of Schuell test profiles.

Ross, G. (1971). 'Discussion of Cormack – A review of classification', *Journal of the Royal Statistical Society Series A* 134: 321-367.

Schuell, H. (1965). *Differential Diagnosis of Aphasia.* Minneapolis: University of Minnesota Press.

Schuell, H. and Sefer, H. (1973). *Differential Diagnosis of Aphasia* (revised). Minneapolis: University of Minnesota Press.

Scott, A.J. and Knott, M. (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 J.* 16: 30-34.

Sneath, P.H.A. (1957). 'The application of computers to Taxonomy', *J. Gen. Microbiol.* 17: 201-226.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy.* San Francisco: W.H. Freeman.

Sokal, R.R. and Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy.* London: Freeman.

Sokal, R.R. and Michener, C.D. (1958). 'A statistical method for evaluating systematic relationships', *Univ. Kansas Sci. Bull.* 38: 1409-1438.

Thorndike, R.L. (1953). 'Who belongs in a family', *Psychometrika* 18: 267-276.

Tyron, R.C. (1939). *Cluster Analysis.* Ann Arbor: Edward Brothers.

Ward, J.H. (1963). 'Hierarchial Grouping to Optimize an objective function', *J. Amer. Statist. Assoc.* 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 J.* 14: 162-165.

Wishart, D. (1969a). 'An algorithm for hierarchial classifications', *Biometrics* 25: 165-170.

Wishart, D. (1969b). 'Mode analysis', pp.282-308 in A.J. Cole (ed.). New York: Academic Press.

Wolfe, J.H. (1965). *A Computer Program for the Maximum Likelihood Analysis of Types.* Technical Bulletin 65-15, U.S. Naval Personnel Research Activity, San Diego.

Wolfe, J.H. (1967). *NORMIX; computational methods for estimating the parameters of multivariate normal mixtures of distributions.* Research Memorandum, S.R.M. 69-17, U.S. Naval Personnel Research Activity, San Diego.

Wolfe, J.H. (1970). 'Pattern clustering by multivariate mixture analysis', *Multiv. 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 distributions. Naval Personnel and Training Research Laboratory (San Diego, California 92152). Technical Bulletin STB 72-2.

Wolfe, J.H. (1978). 'Comparative cluster analysis of patterns of vocational interest', *Multiv. Behav. Res.* 13: 33-44.

Zahn, C.T. (1971). 'Graph-Theoretical Methods for detecting and describing Gestalt Clusters', *IEEE Trans. Computers* C20: 68-86.

Zubin, J. (1938). 'A technique for measuring likemindedness', *J. of Abnormal and Social Psychol.* 33: 508-516.

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.* New York: Human Ecology Fund.