A Review of Classification

By R. M. CORMACK

University of Edinburgh

[Read before the ROYAL STATISTICAL SOCIETY, at a meeting organized by the RESEARCH SECTION on Wednesday, March 10th, 1971, Professor D. J. FINNEY in the Chair]

SUMMARY

The summarization of large quantities of multivariate data by clusters, undefined *a priori*, is increasingly practiced, often irrelevantly and unjustifiably. This paper attempts to survey the burgeoning bibliography, restricting itself to published, freely available, references of known provenance. A plethora of definitions of similarity and of cluster are presented. The principles, but not details of implementation, of the many empirical classification techniques currently in use are discussed, and limitations and shortcomings in their development and practice are pointed out. Methods based on well-defined mathematical formulations of the problem are emphasized, and other ways of summarizing data are suggested as alternatives to classification. The growing tendency to regard numerical taxonomy as a satisfactory alternative to clear thinking is condemned.

1. INTRODUCTION

THE availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other "statistical" innovation (with the possible exception of multiple-regression techniques). The desire to produce a unique labelled pigeon-hole into which an individual entity can be dropped (and then forgotten) is natural to the human brain, or else we have been told this so often that it is now a conditioned response. "A preference for classification is developed in childhood and persists as a habitual form of thought in adulthood" (Goodall, 1954a). Before the new conditioning factor of swelling bibliographies reinforces the reflex we must stop and ask why and when as well as how we should attempt a classification.

A classification, as usually understood, allocates entities to initially undefined classes so that individuals in a class are in some sense close to one another. The process of choosing which of a number of defined classes a new entity should be allotted to is better called *identification* or *assignment* (Dagnelie, 1966). Distinction should be made between three types of classification procedure:

(i) *hierarchical classification*, in which the classes are themselves classified into groups, the process being repeated at different levels to form a tree;
(ii) *partitioning*, in which the classes are mutually exclusive, thus forming a partition of the set of entities;
(iii) *clumping*, in which the classes or clumps can overlap, and a clump and its complement are treated as different types of class.

Distinction must also be made between situations in which entities in one class are or are not required to be distant from entities in another class: situations termed respectively *classification* and *dissection* by Kendall (1966). All collections of entities can be dissected: not all can be classified. "If there are two dense clusters of buildings

separated by much empty space, we have no difficulty in perceiving the existence of two villages; whereas if a village by one name coalesces with a village by another name, we feel that the separation is artificial and that there exist not two entities, but one" (Gengerelli, 1963). In some statistical situations, for example a χ^2 test of goodness of fit to a continuous probability distribution, dissection is the aim. (See Bolshev, 1969, for a summary of the theory of classification in such situations.) In most cases, however, dissection does not serve any purpose.

Terminology is used somewhat haphazardly in the literature. Classification is used to describe the whole subject or it may have either of the restricted uses given above. Cluster methods again refer to the whole subject or to the restricted class of partitioning or clumping. Moreover, one algorithm can lead to different methods of classification: most sorting strategies, for example, lead to a hierarchical classification, which when subjected to a stopping rule gives a partition or clump. I shall not therefore attempt to keep distinct the words "classification" and "clustering"; their meaning should be explicit in each context.

In an amusing and illuminating classification of classifications Good (1965b) lists five purposes:

(i) for mental clarification and communication,
(ii) for discovering new fields of research,
(iii) for planning an organizational structure or machine,
(iv) as a check list,
(v) for fun.

Other authors emphasize (i) and (ii)—"to arrive at a useful description of the sample and to discover unsuspected clusterings which may prove to be important" (Fleiss and Zubin, 1969); "to represent the data in a way which will suggest fruitful hypotheses" (Jardine, 1970); "a classification is predictive with precise purpose unknown at the time of classifying. It cannot be true or false, probable or improbable, only profitable or unprofitable" (Williams and Lance, 1965). However (iv) is expressed explicitly only seldom, and then usually in the context of document retrieval: "to obtain classes such that any member of a class can be treated as if it possessed certain properties", the profile of the class (Jones, 1970)—although this is implicit in any classification. Usually a classification is not intended "to get an answer to a problem which is already set up" (Jardine, 1970) although it is often in practice used, without validity, to test a hypothesis, frequently of the existence of the clusters which it finds.

Classification may be a technique for generating hypotheses: it seems to me that dissection is not. If there are no distinct clusters the data have been forced into a strait-jacket which restricts the domain of possible hypotheses and suggests that some will be generated by the fact of dissection rather than by the data. A hierarchical classification achieves the first four of Good's objectives by providing a concise summary of the inter-relations of the n entities in the form of $(2n-1)$ clusters or the $2n$ links of the tree joining them. Partitioning achieves the first two of Good's objectives by providing a parsimonious summary of the original $n \times p$ data matrix: the n entities are reduced to $n^* \ll n$ classes. Alternative summaries can be achieved by reducing the dimensionality of the variable space from p to $p^* \ll p$ by some *ordination* technique such as principal components or multidimensional scaling. In this argument between classification and ordination the protagonists' viewpoints have been put "on a par with racial prejudices" (McIntosh, 1967). Attempts to reduce both dimensions simultaneously have been made but have not been much used in practical situations.

Three desiderata of a biological classification were laid down by Silvestri and Hill (1964):

(i) objectivity—independent workers should reach similar conclusions;
(ii) stability—the classification should be little affected by new data;
(iii) predictivity—of variates in new individuals.

These need not all always apply to all fields. Sometimes—for example, the classification of an existing complete collection of documents from which information has to be retrieved—there are no new individuals, and (iii) is inappropriate. The stability requirement should also be taken to imply robustness against errors in the data (Jones and Needham, 1968). If a classification is to remain virtually unaltered when extra variables are measured on the same entities, it must clearly be able in some sense to predict unobserved variables. This will be possible only if the unobserved variables are correlated (linear relationship is not necessarily implied) with the observed variables. If a particular observed variable V_1 is the single variable most highly correlated with the other observed variables, it may reasonably be assumed that it is most highly correlated with unobserved variables. This leads, in ordination, to consideration of principal components, and in classification to derived structure methods (Williams and Dale, 1965) not involving any measure of closeness between individuals. We may divide the individuals into clusters according to their value of V_1. This was proposed for binary variables by Williams and Lambert (1959). For such variables such a division may be regarded as a true classification rather than a dissection, by giving infinite weight to V_1 in the distance function used. Division into classes on the basis of the value of a continuous variable V_1 need impose no distance between entities allocated to different classes.

A corollary of this is that, when division between clusters accounts for the maximum correlation between variables, within clusters there should be no correlation between variables. Latent structure analysis (see, for example, Lazarsfeld and Henry, 1968) defines, and seeks, clusters as sets of entities within which the observed variables are independent.

Some writers are of the opinion that classification techniques are now established and the need is for more data: "One of the principal impediments to the development of numerical taxonomy is the difficulty biologists have of measuring and recording taxonomic characters at speeds and in quantities commensurate with the ability of modern computers to process these data" (Sokal and Rohlf, 1966). Others think that "some way is now needed to integrate voluminous new data now being accumulated from all sources into the body of taxonomy" (Sneath, 1969). Unfortunately the current swell of classificatory publications (estimated at more than 1,000 a year) is mainly devoted to "testing" published techniques on data for which "standard" classifications exist. When the technique fails the author's response is to modify the technique instead of thinking about the "standard" classification or questioning the value of the whole process. Let me give one example; not the worst. A well-known technique was applied to specially selected entities typical of a number of established groups. The initial "success" rate of 12 per cent was increased to 60 per cent by a modification of the technique, details of which are not given. Then, by the introduction of features deliberately copying the subjective methods by which the established groups were found, the success rate was increased to nearly 80 per cent. Many questions follow from this finding. What properties of the remaining 20 per cent of the entities cause their misclassification? Are these entities demonstrating a basic failure in the standard

classification? Would other numerical techniques give a "better" result? None of these questions are raised in the paper. The one conclusion, not mentioned by the authors, that I draw from the result is that this numerical technique will be quite useless for classifying an unselected, atypical entity which the standard classification has difficulty in allocating.

2. Measures of Similarity

The basic data can consist of a vector of observations $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ on each entity E_i in a set $\mathscr{E}$, or of a similarity structure S on $\mathscr{E}$ or on $\mathscr{E} \times \mathscr{E}$. In many fields of study these entities are essentially unique individuals. In some, notably in the taxonomy of living organisms, it is assumed that the entities are well-defined populations from which individuals may be sampled. The observations on these individuals have a probability distribution, usually summarized by the mean and covariance matrix. Thus in the classification of such population entities, data on within-entity variation are available in addition to the entity mean observation vector $\mathbf{x}_i'$. In either case, when the co-ordinates of $\mathbf{x}_i'$ are quantitative, binary or ranked (but not unordered qualitative) variables, it is natural to visualize the set $\mathscr{E}$ as n points in $[p]$-space, which, as humans, we should like to be geometric or Euclidean, although this is not a necessary condition.

Hartigan (1967) lists twelve similarity structures:

S1 S defined on $\mathscr{E} \times \mathscr{E}$ is Euclidean distance;
S2 S defined on $\mathscr{E} \times \mathscr{E}$ is a metric;
S3 S defined on $\mathscr{E} \times \mathscr{E}$ is symmetric real-valued;
S4 S defined on $\mathscr{E} \times \mathscr{E}$ is real-valued;
S5 S is a complete order $\leqslant$ on $\mathscr{E} \times \mathscr{E}$;
S6 S is a partial order $\leqslant$ on $\mathscr{E} \times \mathscr{E}$ (each comparable pair of entities can be ordered, but not all pairs of entities need be comparable);
S7 S is a tree τ on $\mathscr{E}$ {a partial similarity order, $(i,j) \leqslant (k,l)$ whenever $\sup_\tau(i,j) \geqslant \sup_\tau(k,l)$} (see Section 4);
S8 S is a complete relative similarity order $\leqslant_i$ on $\mathscr{E}$ for each E_i in $\mathscr{E}$: $j \leqslant_i k$ means that E_j is no more similar to E_i than E_k is;
S9 S is a partial relative similarity order $\leqslant_i$ on $\mathscr{E}$;
S10 S is a similarity dichotomy on $\mathscr{E} \times \mathscr{E}$ in which $\mathscr{E} \times \mathscr{E}$ is divided into a set of similar pairs and a set of dissimilar pairs;
S11 S is a similarity trichotomy on $\mathscr{E} \times \mathscr{E}$ (similar pairs, dissimilar pairs, and the rest);
S12 S is a partition of $\mathscr{E}$ into sets of similar objects.

Most empirical studies have started with one of the structures S1 to S3, and virtually each author has his own details of proceeding from $\mathbf{X}$ to S. Some proposals are listed in Table 1, indices that are measures of dissimilarity, decreasing with increasing similarity, being denoted by $\bar{\text{I}}$. The final three indices are used essentially for binary data and are therefore inapplicable to the classification of population entities. The complement of I8 has been proposed for use with quantitative data—the "Canberra" metric $\bar{\text{I}}8$.

When all variables are quantitative, S1 can be obtained directly, $\bar{\text{I}}1$. Calculation of a distance, however, depends on the scales and inclination of the axes. In the case of unique entities, most workers recommend normalizing each variable by dividing by its standard deviation over the n individuals, although some (e.g. Carmichael *et al.*,

1968) advocate scaling by the range, some by a heterogeneity measure (Hall, 1969a) or a measure of importance (Williams *et al.*, 1964; Hall, 1965), while others (e.g. Sawrey *et al.*, 1960) use the raw data. A comprehensive computer package (e.g. Williams *et al.*, 1966) usually leaves the decision to the user. The unthinking use of

TABLE 1

Indices of similarity

Ī1	Euclidean distance $\sum_{v=1}^{p} w_v(x_{iv}-x_{jv})^2$ Unstandardized: $w_v = 1$ Standardized by S.D.: $w_v = 1/s_v^2$. Denote by Δ^2 Standardized by range: $w_v = 1/\max_{i,j}(x_{iv}-x_{jv})^2$	
Ī2	City-block metric $\sum_{v=1}^{p} w_v \mid x_{iv}-x_{jv} \mid$	Johnson and Wall (1969)
	Mean character difference: $w_v = 1/p$	Cain and Harrison (1958)
Ī3	Minkowski metrics $\left[\sum_{v=1}^{p} \mid x_{iv}-x_{jv} \mid^{1/\lambda}\right]^{\lambda}$	Boyce (1969)
I4	Angular separation $\dfrac{\sum_{v=1}^{p} x_{iv}\,x_{jv}}{\left[\sum_{v=1}^{p} x_{iv}^2 \sum_{v=1}^{p} x_{jv}^2\right]^{\frac{1}{2}}}$	Gower (1967a); Boyce (1969)
I5	Correlation $\rho_{ij} = \dfrac{\sum_{v=1}^{p}(x_{iv}-\bar{x}_i)(x_{jv}-\bar{x}_j)}{\left[\sum_{v=1}^{p}(x_{iv}-\bar{x}_i)^2 \sum_{v=1}^{p}(x_{jv}-\bar{x}_j)^2\right]^{\frac{1}{2}}}$	Sokal and Michener (1958); Fortier and Solomon (1966); McQuitty (1966)
I6	Profile similarity index: $\dfrac{2k_m-\Delta^2}{2k_m+\Delta^2}$, where $P(\chi_p^2 < k_m) = 0{\cdot}5$	Cattell (1949)
I7	Coefficient of nearness: $\{\sqrt{(2p)}-\Delta\}/\{\sqrt{(2p)}+\Delta\}$	Cattell and Coulter (1966)
Ī8	"Canberra" metric: $\sum_{v=1}^{p} \mid x_{iv}-x_{jv} \mid/(x_{iv}+x_{jv})$	Bray and Curtis (1957); Lance and Williams (1966)
I8	$\dfrac{2a}{2a+b+c}$	Czekanowski (1913); Dice (1945)
I9	$\dfrac{a}{a+b+c}$	Jaccard (1901); termed "connection" by Needham (1963)
I10	Simple matching: $\dfrac{a+d}{a+b+c+d}$	Sokal and Michener (1958)

scaling must be condemned, even more in this context than in standard multivariate analyses. Not only is there the same argument that the difference in scale between two variables (particularly of the same dimension) may be intrinsic, but also, as Fleiss and Zubin (1969) have pointed out, the scaling should be carried out on individual clusters while it is in fact carried out on the complete set of data.

Similar considerations affect the problem of correlated variables. Sokal (1961) pointed out that $\sum(x_{iv}-x_{jv})^2$ is not Euclidean distance in this case. If the entities to be classified are populations from which several individuals can be sampled, information is available on the scales of, and the correlation between, the variables within these populations. Thus Mahalanobis's D^2, as advocated by Bolshev (1969) and, in a form modified for discrete data, by Balakrishnan and Sanghvi (1968) and by Kurczynski (1970), can be used. If the within-entity covariance matrices are reasonably similar, a pooled matrix can be used to determine the appropriate axes with reference to which distances between entities should be evaluated.

The matrix of covariances between variables calculated from n unique entities has been used to transform the axes of measurement. Minkoff (1965) found that this yielded results less in agreement with expectations than the correct use of D^2. There are two major objections to basing a distance measure on such an overall covariance matrix. Firstly, most of the correlation present is likely to be caused by the existence of the clusters being sought. This must be retained (Gower, 1969b). Secondly, the correlation structure within clusters may vary considerably from cluster to cluster, so that a pooled covariance matrix is inappropriate. The first point does not wholly imply that "in general, the contribution of any one property-resemblance to overall similarity should not be influenced with respect to other features" (Hall, 1969b), unless care has been taken to ensure that all variables measured are in fact uncorrelated within the clusters.

If replicates do not exist, there appears to be a circularity in trying to transform data by properties of the cluster that it is hoped to determine. Rohlf (1970), however, has proposed a sequential scheme for cluster formation in which distances from an already clustered individual are measured in the local geometry of that cluster. Such a distance is not symmetric—type S4. This procedure eliminates both scaling and correlation problems, by being invariant under any linear transformation of the original variables. A different way round these problems is found in Gower's (1966) proposal to replace the standardized observed variables by principal components before calculating Euclidean distances.

Euclidean distance also has the property, disliked by some users, of giving extra weight to outlying values of a single variate. This is partly overcome by scaling. However, some sets of variables (e.g. measurements of different plant species in an area) seem unsuited to scaling, and it is possible (Bannister, 1968) for two areas containing identical botanical species in differing amounts to be further apart than two areas with no species in common. Some objections to Euclidean distance reduce to the complaint that it does not behave in the desired way: "taxonomic distance has distortions that make it clearly not suitable" (Hall, 1969), although distortion is not defined.

Similarity indices with properties akin to correlation coefficients are often sought. Cattell has advocated a series of such indices of which I6, I7 are examples. The correlation coefficient I5 is not often used. Some arguments against it are circular, amounting to saying that it can give $\rho_{ij} < \rho_{ik}$ when entities E_i, E_j are obviously more similar than E_i, E_k (Eades, 1965). But use of the correlation coefficient must be restricted to situations in which variables are uncoded, comparable measurements or counts; it is not invariant under scaling of variables, or even under alterations in the direction of coding of some variables (Minkoff, 1965).

If the variables are all binary characters, all coefficients of association from the 2×2 table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ of numbers of characters possessed or not possessed by entities E_i, E_j

are candidates for an index of similarity. The properties of these S3 measures are discussed thoroughly by Sokal and Sneath (1963) and Dagnelie (1965). The choice among I8, I9, I10 is dependent on whether co-absence of a particular character is assumed to contain information. Rayner (1966) distinguishes between *dichotomies* (*d* unimportant) and *alternatives*. Intermediate weightings of *a*, *d* can be used. Such a scheme was used by Hayhoe *et al.* (1964) in order to incorporate in the similarity index the differing frequencies of occurrence of different attributes, co-presence of a rare attribute scoring much higher than its co-absence or than the co-presence of a common attribute. This approach has been condemned by some on the basis that it is unjustified scaling. The simple matching coefficient I10 was supported by Williams and Dale (1965) as being the one-complement of a Euclidean distance, while I8 and I9 are not, although their one-complements do satisfy the triangle inequality (Ihm, 1965). Versions of I8, I9, I10 have been used with quantitative variables, I8 most extensively.

The probabilistic distribution of the simple matching coefficient was evaluated by Goodall (1967) on the assumption of independent attributes. This assumption, which may be reasonable for established clusters though not for the primal mix, also underlies his probabilistic similarity index (1964, 1966a). For each attribute separately a similarity index for each pair of entities is defined as the probability P_u that a random pair will have similarity strictly less than the pair in question. The overall similarity of the pair is based on $-2\sum_{u=1}^{p} \ln P_u$ and appeal is made to the χ^2_{2p} distribution to find s_{ij}. Goodall shows empirically that for random data the distribution of s_{ij} closely approximates a uniform distribution.

Difficulties arise with binary data if a character, when present, can be in one of a number of secondary states. A flower with petals can have white, red or striped petals but a flower without petals is just that. Kendrick and Proctor (1964) argued strongly for weighting primary characters by one more than the number of possible secondary characters (*m*). Williams (1969) has re-examined this and shows that $1/m$ is a more suitable weighting for secondary mis-matches. Other authors proposed that the absence of a primary character in such circumstances be recorded as a "no comparison", that variable to be excluded from that entity's similarities. Beers and Lockhart (1962) and others made use of this idea to incorporate variables with more than two ordered states by replacing each such variable by a string of binary ones. The most comprehensive treatment of such similarity coefficients is provided by Rubin's (1967) modified fractional match coefficient which makes allowance for non-applicability of variables and for inclusion of some continuous variables as fractional matches. The same procedure was proposed by Rayner (1966). Rubin argues strongly against discretizing continuous variables, as had been proposed by Rogers and Tanimoto (1960) among others, and warns of the dangers of many-state variables whose effect on s_{ij} will be swamped by a binary variable. Some guide to the magnitude of the effect is given by Cochran and Hopkins (1961).

A similar type of problem arises when it is not clear what variable in one entity corresponds to a particular variable in another. A sound mathematical basis for assessing homologies was formulated by Jardine (1967), who showed that two characters are homologous if their similarity with respect to spatial relation was strictly greater than that of either character in the one individual with any other character in the other. Rayner (1966) had adopted heuristically a similar approach to soil horizons, although his proposals have been disputed, for example by Grigal and Arneman (1969). The effect of making an error in assessing homologues has been explored by Fisher and Rohlf (1969) by the numerical device of randomizing the order

of some variables independently on each entity. In their example, 6 scrambled variables out of 74 affected the final classification little but with 10 the structure of the classification was much less clear.

Any monotonic function of a similarity index s_{ij} can also act as a similarity (or dissimilarity) function (see, for example, de la Vega, 1967). Choice of function depends on the representation sought for the data. In an attempt to develop the intuitive analogy between similarity and information, Rogers and Tanimoto (1960) used $[-\log_2(s)]$. The attempt fails for two reasons: the act of taking logarithms does not in itself turn similarity into entropy, and the logarithms are not additive if the variables are dependent (Lance and Williams, 1967b). Ihm's (1965) further criticism that $[-\log_2(s)]$ is not a metric seems to misunderstand the reason for the transformation. In exploring the representation of multivariate data as entity-points in a Euclidean space, Gower (1966, 1967a) proposes two further transformations. He shows that $d = \sqrt{\{2(1-s)\}}$ can function as Euclidean distance provided that S is positive semi-definite. When the variables are ranked in order for each entity the Euclidean representation of the entity points fall on a hypersphere, in which case the angular separation $[\cos^{-1} s]$ is a possible alternative transformation. The first proposal has been extended by Gower (1967b) to demonstrate the geometrical implications of various clustering techniques, and can be developed to provide a similarity index suitable for mixed qualitative and quantitative data. In this a similarity is defined for each attribute:

$$\begin{aligned} s_{iju} &= 0 \quad \text{if attribute } u \text{ qualitative,} \quad x_{iu} \neq x_{ju}, \\ &= 1 \quad \text{if attribute } u \text{ qualitative,} \quad x_{iu} = x_{ju}, \\ &= x_{iu}x_{ju}/f_u \quad \text{if attribute } u \text{ quantitative on scale } f_u. \end{aligned}$$

Overall an index of similarity is $s_{ij} = \sum_{u=1}^{p} s_{iju}/p$ which can be defined recursively over attributes. Since s_{iik} is not necessarily 1, $d_{ij} = \sqrt{(s_{ii}+s_{jj}-2s_{ij})}$ must be used for a distance measure. The geometrical implications of this d_{ij} will be considered more fully later. An alternative index of similarity for mixed data uses a quantified version of I8 together with the one complement of $\bar{\text{I}}8$ (Lance and Williams, 1966).

Standard statistical "measures" of association have not been overlooked, ecologists in particular having made considerable use of functions of χ^2 from the 2×2 table of presence and absence. This implies that co-absence of any character is as important as co-presence. Williams and Lambert (1959) used both χ^2 itself and $\sqrt{(\chi^2/p)}$, the latter being identical with the correlation coefficient. Although these were used in a derived structure analysis evaluating the association between variables, they have been used to measure similarity between entities despite the dangers of their dependence on p, which makes χ^2 an unsuitable measure, albeit a good test of association. Cole (1949) introduced a series of such measures, and a discussion of the features required in such a coefficient by ecologists can be found there or in Hurlbert (1969). Hurlbert adds another measure

$$\frac{ad-bc}{|ad-bc|}\sqrt{\left(\frac{\text{obs}\,\chi^2 - \min\chi^2}{\max\chi^2 - \min\chi^2}\right)}$$

to the list, claiming it as "the most satisfactory measure of association for use with presence–absence data". Comprehensive bibliographies of earlier measures are given by Goodman and Kruskal (1959), Dagnelie (1960) and Sokal and Sneath (1963).

The idea that any monotonic transformation of s_{ij} also provides a possible similarity index has led to various proposals for simplifying the very large matrices that can be involved. The crudest (S10) chooses a threshold t and sets $s^*_{ij} = 1$ if $s_{ij} \geqslant t$, $s^*_{ij} = 0$ if $s_{ij} < t$ (Bonner, 1964; Estabrook, 1966; Parker-Rhodes and Jackson, 1969). The similarities can then be represented as a graph on $\mathscr{E}$. Less prodigal of information is the approach, worked out extensively by Lerman (1969, 1970), which takes as informative data the ordering induced by s on $\mathscr{E} \times \mathscr{E}$ (S5 and S6). He takes as basic definition of a similarity measure any function on the 2×2 table (a, b, c, d) which is increasing in a, symmetrical in b and c, and decreasing in d. Using straightforward probability arguments, he deduces that:

(i) the two similarity measures whose ordering relations are most unlike are a and $(a+d)$. They are unlike in the sense that the cardinal $|\Delta|$ of the symmetric difference Δ between the graphs of the ordering on $\mathscr{E} \times \mathscr{E}$ induced by the two measures is maximum;

(ii) if all individuals possess either m or $(m+1)$ attributes, for some m, $0 < m < p$, the orderings induced by any two similarity measures are identical;

(iii) if the variance over the individuals of the number of attributes possessed by any individual is V, then $|\Delta| < \frac{1}{9}n^2(n-1)^2 V$ for any two similarity measures.

The extension of these arguments to the assessment of clustering criteria will be discussed later.

3. Clustering Techniques

There are many intuitive ideas, often conflicting, of what constitutes a cluster, but few formal definitions. Two basic ideas are involved: internal cohesion and external isolation. Sometimes isolation is stressed: Rogers *et al.* (1967) found the maximal acceptable restriction to be that similar entities shall not be placed in different classes and that a discontinuity should be observable between classes. Sometimes cohesion is stressed: an individual should be accepted into a cluster if its smallest correlation with any member is greater than some threshold (Cattell, 1944). More usually, both are included: the distance between any two points in the set is less than the distance between any point in the set and any not in it (Gengerelli, 1963); the sum of the similarities of any member to the other members should exceed the sum of its similarities to non-members and vice versa for non-members (Needham, 1963).

In the social sciences the search has been for tight clusters or cliques in which each entity resembles every other, and in which all are satisfactorily described by one—the profile of the set. Often these have turned out to be will-o'-the wisps. Even when found they are often seen to be not necessarily unique (Cattell and Coulter, 1966), as one entity can be a member of more than one clique. Cliques thus proved difficult to define (Fisher, 1969) even when present, and most subjects have accepted the need for a more general idea of group. Coleman and MacRae (1960), for example, found that "in large measure they (i.e. groups) are composed of chains of choices and have an octopus-like configuration rather than a clear division into cliques".

A weak definition of cluster will allow such multidimensional amoebae (Needham and Jones, 1964). Are they acceptable? Without a formal definition anything can be debated. The general aim should be to describe the data in a simpler way than the original (Jones, 1970) without incurring offensive mathematical consequences (Needham, 1965a). Some find amoeboid structures unacceptable, others reject any unnecessary restriction to spherical clusters (Needham, 1965a; Rogers *et al.*, 1967),

although the most commonly used techniques impose severe variance restraints (Wishart, 1969c) which are bound to result in hyperfootballs (Needham, 1965b). The only attempt to set down universally acceptable mathematical criteria for deciding whether a set $\mathscr{E}$ should be partitioned was made by Rubin (1967). There has been no comment by users of clustering techniques as to whether these criteria are acceptable.

To provide a basis for discussion, Jones and Jackson (1970) have recently given labels to certain forms of connected structure of sets $\mathscr{E}$ in which each pair E_i, E_j either is or is not linked (S10). Their labels do not seem to have any advantages over m-cliques, introduced by Luce (1950). An m-clique is a subset of entities of which every entity is connected to every other entity by a chain of not more than m links, one pair at least requiring the full m links: this subset cannot be a proper subset of another set with the same properties. A recent proposal by Jardine and Sibson (1968a) for k-partitions of $\mathscr{E}$ forms a kind of dual to m-cliques: their suggestion will be discussed later.

Most techniques for clustering have developed, without formal basis, as algorithms. A formal approach would set up a criterion to be optimized over the set of partitions of $\mathscr{E}$. Unfortunately there are far too many partitions of n, for $n > 20$ say, for a complete enumeration to be feasible. The search must be conducted over a limited range of partitions. Three types of procedure are in general use for finding clusters:

(a) agglomerative—a series of successive fusions of the n entities into groups;
(b) divisive—partition of complete set $\mathscr{E}$ successively into finer partitions;
(c) clustering—successive re-allocation of individuals between the sets of some initial partition.

Of these, (a) and (b) are methods for representing the data as a dendrogram (see Section 4), from which clusters are obtained by cutting at any level; (c) are procedures for finding directly a partition of $\mathscr{E}$ with properties approximating to some desiderata.

Some sorting strategies do yield clusters with well-defined properties and hence are exact algorithms for a properly defined method. Williams and his co-workers have often made the distinction between the clusters and the route by which the clusters are obtained, but even an exact algorithm for a properly defined method is not necessarily optimal. With other sorting strategies, the resulting clusters are defined only by the algorithm by which they were obtained. Unless the clusters can be shown to have properties approximating to some desiderata *for the clusters*, the fact that they have been obtained by successive steps each of which was best of the steps then available seems irrelevant. Like least-squares estimation of a non-linear structure, such procedures give results that are best of an undefined, and possibly undesirable, class. But the analogy goes further: the algorithms can be carried through and an answer obtained, whereas other better defined methods can perhaps not be implemented.

Inter-cluster similarity: a selection of definitions of similarity between entities is available. The similarity between clusters must also be defined. The measures proposed mostly satisfy a recurrence formula for the dissimilarity between group k and a group (ij) formed by the fusion of groups i and j (Lance and Williams, 1966b, 1967a, Anderson 1971a):

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|.$$

The values of the parameters for several well-known sorting strategies are given in Table 2. If similarities s_{ij} rather than dissimilarities d_{ij} are given, the same relation holds, with d_{ij} replaced by $s_{ij} = 1 - d_{ij}$; if $\alpha_1 + \alpha_2 + \beta = 1$, s can replace d directly.

These relations are usually applied in clustering as an agglomerative procedure, starting with a similarity or dissimilarity matrix between entities. A sorting procedure not satisfying a relation such as the above has the computational disadvantage of requiring the initial data, in addition to the cluster data, to be retained at all stages.

TABLE 2

Sorting strategies

	Name	α_i	β	γ	*References*
L1	Single linkage (nearest neighbour)	$\frac{1}{2}$	0	$-\frac{1}{2}$	Sokal and Sneath (1963)
L2	Complete linkage (furthest neighbour)	$\frac{1}{2}$	0	$\frac{1}{2}$	Sokal and Sneath (1963); McQuitty (1964)
L3	Group average	$n_i/(n_i+n_j)$	0	0	Sokal and Michener (1958); McQuitty (1964)
L4	Weighted average	$\frac{1}{2}$	0	0	McQuitty (1966, 1967a)
L5	Centroid	$n_i/(n_i+n_j)$	$-n_i\,n_j/(n_i+n_j)^2$	0	Sokal and Michener (1958); Gower (1967b)
L6	Median	$\frac{1}{2}$	$-\frac{1}{4}$	0	Gower (1967b)
L7	Minimum variance	$(n_i+n_k)/(n_i+n_j+n_k)$	$-n_k/(n_i+n_j+n_k)$	0	Wishart (1969b); Anderson (1971a)
L8	Flexible	$\frac{1}{2}(1-x)$	$x(<1)$	0	Lance and Williams (1967a)

Standard arithmetic procedures, however, are not invariant under monotonic transformations of s. If this is required attention is restricted to L1, $\min(d_{ki}, d_{kj})$, and L2, $\max(d_{ki}, d_{kj})$ (Johnson, 1967).

Nearest neighbour is the simplest agglomerative sorting procedure and requires only a similarity structure in the form of a complete ordering on $\mathscr{E} \times \mathscr{E}$. The procedure may be continued to give a complete dendrogram, in which form results are usually expressed, or may be terminated at any threshold t. The clusters so formed are defined by the condition that two entities E_i, E_j belong to the same cluster if there exists a chain of entities $E_k, E_l, \ldots, E_q, E_r$ such that $s_{ik}, s_{kl}, \ldots, s_{qr}, s_{rj}$ are all greater than t. Sorting strategy L1 is an exact algorithm for such single-linkage clusters. Entities tend to be incorporated into existing clusters rather than joined to form the core of an independent cluster. Williams *et al.* (1966) define a coefficient of chaining to give numerical expression to this tendency. In the absence of a formal definition of chaining such a coefficient gives only spurious precision to any argument about the undesirability of this property. Lance and Williams (1967a) term this feature of the sorting strategy "space-contraction", a concept whose value is reduced by its lack of formal definition. One advantage of single linkage is that successive fusions always occur at lower levels of inter-cluster similarity.

Single linkage does not give satisfactory results if intermediates are present between clusters (Hodson *et al.*, 1966). Such intermediates can be viewed as the

result of random noise, and methods have been proposed for eliminating noisy entities (Wishart, 1969) or noisy variables (Baron and Fraser, 1968). Shepherd and Willmott (1968) suggest imposing a requirement that an entity is added to a cluster only if its similarities to k or more members of the cluster are all greater than some threshold.

Furthest neighbour sorting also requires only a similarity ordering on $\mathscr{E} \times \mathscr{E}$. If the dendrogram is terminated at a threshold t the clusters so formed are defined by the condition that the similarity between all pairs of entities in a cluster must be greater than t. Strategy L2 is an exact algorithm for such complete-linkage clusters. Since the similarity of a single entity with a cluster is the minimum of its similarities to the individual entities in the cluster, this strategy produces compact clusters with no chaining. Again successive fusions occur monotonically with inter-cluster similarity.

Group average methods require numerical similarity indices and are intermediate in effect between L1 and L2. Only similarity indices whose average is meaningful should be used. The correlation coefficient should not be chosen (Lance and Williams, 1967a) unless first turned into a covariance (King, 1967). The similarity between groups is given by

$$\sum_{i \in A} \sum_{j \in B} s_{ij} w_i w_j / \sum\sum w_i w_j$$

(Sokal and Sneath, 1963). Strategy L3 takes $w_i = 1$, L4 takes $w_i = n_A$. These strategies are monotonic. However, the clusters formed are defined only by the strategies.

Centroid sorting has its origins in the characterization of the data matrix as points in Euclidean space. Every cluster is regarded as a single point at its centroid. Agglomerative techniques fuse either the two clusters with minimum between-centroid distance (L5) or those which yield minimum within-cluster variance (L7). However, the former is not monotonic.

Although the geometrical basis of the method suggests that Euclidean distance $\bar{1}1$ be used as dissimilarity index, the method can be used with any index, although a recurrence relation such as those in Table 2 need not result. The method first appeared as Sokal and Michener's (1958) unweighted pair-group method. The geometric properties are given by Gower (1967b).

With $\bar{1}1$ as dissimilarity measure, it is natural to seek to minimize the within-group S.S. at each fusion (L7). The k-partition of $\mathscr{E}$ resulting at any level of the dendrogram may be regarded as an approximation to the k-partition minimizing the total within-group S.S. This is a measure of the disorder in the system. Algorithms and discussions have been given by Ward (1963), Ward and Hook (1963), Orloci (1967) and Wishart (1969b). A related method, described earlier by Sawrey *et al.* (1960), built up a hierarchy from mutually dissimilar nucleus groups by incorporating in each group those points nearer than a preassigned threshold: centroids and distances were recomputed for these and the procedure repeated at a lower threshold.

In one dimension only contiguous partitions need be considered. For small problems Fisher (1958) proposed finding the true minimum by enumeration, for larger ones approximating the minimum by hierarchical splitting, a technique adopted for small multidimensional populations by Edwards and Cavalli Sforza (1965). The true significance levels for Fisher's one-dimensional problem have been obtained by Engelman and Hartigan (1969). The relationship between clustering and multiple comparison tests has been explored by Calinski (1969): although overlapping clusters are usually apposite, Engelman and Hartigan's table might yield instructive comparisons with standard procedures.

Fisher (1969) considers a generalization of the problem to allow weighting of the points and their interactions so that there is a cost involved in expressing the n points as n^* centroids. The aim is to minimize this cost. He allows a general metric similarity which can be transformed to Euclidean distance in the way proposed by Gower (1966). The algorithm improves on that of Ward (1963) by trying some sub-optimal fusions at specified places in the hierarchy in the hope that one of these may lead to a better clustering than that obtained by optimizing the hierarchical route. Fisher suggests that, for small n, the degree of optimality attained be assessed by continuing the fusion until there are only two clusters, whose composition can be compared with the optimal found from enumerating all such partitions. Some examples are moderately encouraging.

Information theory

If a set $\mathscr{C}$ of c entities, characterized by p attributes, contains a_u entities possessing the uth attribute, the information content of the set can, following Shannon, be defined as

$$I(\mathscr{C}) = pc\log c - \sum_{u=1}^{p}\{a_u \log a_u + (c-a_u)\log(c-a_u)\}.$$

$I(\mathscr{C})$ is a measure of the disorder of the set $\mathscr{C}$, so that centroid agglomerative sorting to minimize $\sum I(\mathscr{C}_i)$ can proceed as in the preceding section (Williams *et al.*, 1966). Either the gain in information ΔI can be minimized or the sets fused whose fusion has minimum I (Lambert and Williams, 1966). Since I is strictly additive the hierarchy is certainly monotonic in I, and appears also to be monotonic in ΔI. The significance of any fusion can be assessed by appeal to the χ^2 approximation for $2\Delta I$ or for I; Field (1969) has warned that the degrees of freedom should be one less than the number of attributes present in the sets under study. The strategy strongly favours clusters of equal size (Lance and Williams, 1966a). The data are being forced into a particular structure without any formal examination of that structure. The possibility of using other information theory models for both binary and frequency data are discussed by Orloci (1969). The relation between information and likelihood has brought information concepts into several classification studies (Rogers and Tanimoto, 1960; Macnaughton-Smith, 1965; Harrison, 1968). Sneath (1969a) has pointed out that entropy is not identical with clustering tendency since a regular distribution also has low entropy. A full discussion is given by Theil (1967). The clusters obtained from information measures have been found to be very sensitive to the actual set of entities included (Hall, 1967; Gower, 1969a). The early uses of information in classification were due to Rescigno and Maccacaro (1961).

An ambitious attempt to express in information terms the complete classification process from initial data to labelled classes has recently been made by Boulton and Wallace (Wallace and Boulton, 1968; Boulton and Wallace, 1970). Variables are assumed independent within classes, but both continuous (here assumed normal) and multinomial variables can be incorporated. Replicate entities are required.

Iterative relocation

When a criterion to be minimized is well defined, as in the last two sections, a direct attempt to find a partition that minimizes it, rather than the indirect hierarchical approach of the preceding sections, can be made. Both procedures are sub-optimal, but in different respects.

Initially k points are chosen as cluster centres: either randomly (MacQueen, 1967), regularly spaced (Beale, 1969a, b), mutually farthest apart (Thorndike, 1953: cf. Kennard and Stone, 1969) or suitably chosen either from the data (Ball and Hall, 1967; Nagy, 1969) or supplementary to the data (Jancey, 1966). All points are then allocated to the nearest centre. The centres may be changed after each point has been allocated (e.g. MacQueen, 1967) or only after all points have been allocated (e.g. Ball and Hall, 1967), to the centroid of the new cluster. Nagy (1969) gives a modification which permits overlapping clusters. MacQueen (1967) establishes that his procedure has satisfactory asymptotic properties.

Beale proposes a first trial with k larger than should be necessary. When the iterations stabilize the pair of clusters which increases least the total S.S. are merged and the process repeated. On the assumption that clusters are spherical normal the contribution of extra clusters in reducing the residual S.S. can be tested by an F-statistic.

Both MacQueen (1967) and Ball and Hall (1967) by slightly different mechanisms allow k to change during the relocations by enforcing the division of a cluster with large within-cluster S.S. or the fusion of two clusters with small between-cluster S.S. according to two parameters of coarsening and refinement set by the user. MacQueen notes a tendency for final clusters to be of comparable size. He proposes to test the clustering by computing the within-class variance when the values in each dimension are randomly associated, and comparing the observed minimum variance with the randomization distribution formed by repetitions of this process.

Intuitively one would feel that using as initial allocation the sub-optimal result obtained from a hierarchical structure would lead to more rapid convergence; this has been found by Nagy (1969). However, in a series of trials of a general relocation procedure, Wishart (1971) found that the procedure converged more rapidly, and to the optimal solution, from an extremely bad initial value than from a nearly optimal one.

An alternative criterion to be minimized by iterative relocation has been proposed by Rubin (1967). He defines the stability of an entity E_i in a cluster C of n_c entities as

$$\frac{\alpha}{n_c - 1} \sum_{j \in C} s_{ij} - (1-\alpha) \max_{j \notin C} s_{ij}.$$

An interesting innovation in Rubin's approach is that he allows new groups to form by considering an empty set whose similarity with every entity is α. From an initial partition we try to maximize the average entity stability. Another innovation worthy of examination is the provision of a residue set of unclassifiable entities, those entities whose stability in the optimal partition is negative.

The concept of relocation of entities gives these techniques an immense advantage over agglomerative sorting (Lance and Williams, 1967b).

Divisive techniques

The number of ways of partitioning a set of n entities into m groups is too immense for all to be examined even for $m = 2$ (see, for example, Fortier and Solomon, 1966):

$$P(n,m) = \left\{ m^n - \sum_{i=1}^{m-1} m_{(m-i)} P(n,i) \right\} \Big/ m!$$

Monothetic divisions on the presence or absence of a single suitably chosen attribute are feasible, but, since they do not depend on the similarities between entities, will be

discussed later. The only feasible suggestion for a polythetic (i.e. dependent on several attributes) divisive technique is dissimilarity analysis (Macnaughton-Smith *et al.*, 1964). In this a splinter group $\mathscr{A}$ is accumulated by sequential addition of the entity whose total dissimilarity with $(\mathscr{E}-\mathscr{A})$ less its total dissimilarity with $\mathscr{A}$ is maximum. When this difference becomes negative the process is repeated on the two sub-groups. As with all divisive schemes an inefficient early partition cannot be corrected later. The advantage that Macnaughton-Smith (1965) claims for divisive techniques "that statistical error is more troublesome when we start at the 'bottom of the page'" seems totally unjustified for a set of entities for which no common error structure is assumed.

Single-cluster formation

An alternative strategy to building the whole dendrogram simultaneously is to complete each cluster that is initiated before starting a new cluster. This technique has been used by McQuitty (1964) and Carmichael *et al.* (1968) using L1, and by Sokal and Michener (1958), Kendall (1965) and Hope (1969a) using L3. Hope places the extra requirement that inter-cluster fusions up to a chosen threshold be made before entity-cluster ones. He forms a dendrogram by a set of decreasing thresholds. Otherwise an arbitrary stopping rule has to be chosen, usually on the basis of some discontinuity in the similarity with the cluster of the next entity to be incorporated. In this case, overlapping clusters are a logically inescapable development. McQuitty's proposed stopping rule—to add a new entity only if it is nearer to a point in the cluster than to a point not in the cluster—seems less arbitrary, avoids overlapping clusters and undoubtedly lessens chaining.

Comparative studies

If clusters are really distinct it would be hoped that any strategy worthy of use would find them (Gower, 1969b). Many studies have found close agreement between different strategies (Sneath, 1966; Watson *et al.*, 1966; El Gazzar *et al.*, 1968; Muir *et al.*, 1970) and even between different subsets of variables (Grigal and Arneman, 1969). Other authors, however, report much less consistent results. For example, Sammon (1969) finds the results of clustering procedures highly dependent on the choice of index, thresholds and number of re-allocation iterations. See also papers by Minkoff (1965) and Colman (1968).

Two types of study seem likely to be highly informative but have been surprisingly little carried out. Firstly: studies of real data if either the variables or the entities are divided into subsets, either randomly or to a chosen pattern, and clustering performed independently on each subset. The example given by Lange *et al.* (1965) is a devastating comment on the results of Williams and Lambert's (1959, 1960) association analysis. Single linkage fared much better.

Secondly: studies of concocted data of known structure. Wishart (1971) has used an iterative relocation scheme to compare different similarity indices in a manner unaffected by sorting strategy. From their inability to separate correctly four well-separated bivariate normal samples, he recommends that the correlation coefficient I5 and certain other indices not discussed here be debarred from further use.

Such empirical studies provide critical information of the behaviour of different similarity measures and sorting techniques in practice. Theoretical studies are less easy, but may be valuable. Lerman (1969) has studied two criteria for assessing a set of clusters based on the sets $\mathscr{S}, \mathscr{D}$ of pairs of entities in the same or different clusters:

(a) the number of times a pair in $\mathscr{D}$ are less similar than a pair in $\mathscr{S}$;
(b) the excess of (a) over the number of times a pair in $\mathscr{D}$ are not less similar than a pair in $\mathscr{S}$ (de la Vega, 1967).

By a combinatorial argument the first of these is shown to be unsatisfactory in that, whatever the original similarities so long as no two are equal, certain partitions cannot maximize (a). For example 49 objects can never be partitioned into 7 sets of 7. Tendencies for criteria to give preference to certain patterns, such as discussed by Gower (1967b), can be tolerated if understood: absolute exclusion of some patterns renders a criterion wholly unacceptable.

Lerman does not comment on the effect of expressing the criteria as proportions of the number of $(\mathscr{D}, \mathscr{S})$ pairs of pairs. In this form they are essentially equivalent, and (a) is numerically identical to a measure $g(-,0)$ proposed by Jackson (1969) for the "discrepancy" between two similarity matrices ("discrepancy" appears to be maximal for identical matrices). Jackson suggests that more informative results may come from considering for different thresholds t only pairs (i,j) such that $s_{ij} > t$.

Miscellaneous probabilistic considerations

Various proposals have been made to take account of random variation in cluster formation. McQuitty (1956) notes that some matches in a simple match coefficient will occur by chance, and proposes subtracting the expected number of such matches from the observed number before calculating the similarity index. He defines a group's attributes as those that are common to all members of the group so that chance matches between groups are fewer than between individual entities.

Goodall (1968) extends his probabilistic similarity index to evaluate the affinity of an entity to an existing cluster. For each attribute in turn he arranged all its possible values in order of their similarity to the cluster norm, defined differently for different types of attribute. The tail probability of an entity being further from the cluster than it is can be estimated from the relative frequency of such events, and these probabilities combined as before over attributes to give a χ^2_{2p} variable. Goodall (1966b) has argued strongly that the hypothesis of a single cluster should be tested.

With a fixed set $\mathscr{E}$ of entities the idea, implicit in this and in MacQueen's test of clustering, of a null hypothesis that the set of variable values observed are randomly allocated to entities seems a proper and the only possible approach. If $\mathscr{E}$ is regarded as a random sample of entities from an infinite population then models such as spherical normal distributions can be invoked, and the standard type of significance test applied. Two difficulties loom large: the formulation of appropriate null hypotheses and the evaluation of the probability distributions of the maxima that will be used as test statistics. With a fixed set of entities any discontinuity is important whether statistically significant or not. In deciding how, or whether, to cluster a sampled set of entities such statistical problems must be faced. An example of a possible approach has been given by Goodall (1966a).

If the entities are populations, for each of which a particular probability distribution of variables may be assumed, the problem of allocating a new individual to an existing entity is the statistical one of multiple discrimination. The problem of clustering, or forming a hierarchy, of the entities differs from that of clustering unique entities only in so far as estimates of covariance matrices within entities permit more effective probabilistic indices of similarity to be used.

4. Representations of Similarity Structures

Several authors have written down the logical stages in handling a large body of multidimensional data. In Jardine's (1970) framework the process should be:

(1) Set up a precise mathematical characterization of the data and of the representation wanted.
(2) Lay down criteria of adequacy for transformations from the data to the representation (e.g. invariance, structure, optimality).
(3) Examine the transformations for existence, uniqueness and further properties.
(4) If these are non-constructive or non-feasible, seek efficient algorithms for implementing them.

Features of the representation will be denoted by *.

Hierarchical structures

Hartigan's (1967) similarity structures, discussed earlier, form one set of characterizations of the data. Hierarchical stratified clustering (S7) is the most carefully defined representation, given independently by Constantinescu (1966), Hartigan (1967), Jardine *et al.* (1967), Johnson (1967) and Lerman (1970). Formally it is a tree $r = [\mathscr{E}, \mathscr{C}, T]$ consisting of a root $\mathscr{E}$ (the cluster containing all entities), a finite set $\mathscr{C}$ of nodes C (clusters of entities) and a mapping T of $\mathscr{C}$ into itself such that for all $k \geqslant 1$, $T^k C = C$ if and only if $C = \mathscr{E}$, together with a real valued function σ on $\mathscr{C}$ such that $\sigma(C) \leqslant \sigma(C')$ if there exists a $k \geqslant 0$ such that $T^k C' = C$. The pictorial representation as a dendrogram has long been used to describe classifications. The similarity s_{ij}^* between two clusters or two entities is defined as the value of σ at the first node in which they are united by T. The dissimilarity $d_{ij}^* = 1 - s_{ij}^*$ is an ultrametric satisfying $d_{ij}^* \leqslant \max [d_{ik}^*, d_{jk}^*]$ for all i, j, k.

The representation distorts the characterization. When the characterization is numerical, various measures of distortion have been proposed and are given in Table 3. When the characterization is an ordering on $\mathscr{E} \times \mathscr{E}$, two further measures have been proposed by Lerman (1970). These depend on consideration of sets of three entities E_a, E_b, E_c in $\mathscr{E}$. If $s_{ab} \leqslant s_{bc} \leqslant s_{ac}$ the triple has ultrametric structure if and only if $s_{ab} = s_{bc}$. If $s_{ab} < s_{bc}$ there will be $g(a, b, c) \geqslant 0$ pairs of entities whose mutual similarity lies in the open interval (s_{ab}, s_{bc}). Lerman proposes either the average or the maximum of g over all triples in $\mathscr{E} \times \mathscr{E} \times \mathscr{E}$ as a suitable distortion measure. Roux (1969) independently considers the same representation and gives a systematic procedure for making all triples ultrametric.

Jardine *et al.* (1967) also make precise what many authors have suggested, namely that the transformation from **S** to **S*** should be (a) well defined and (b) continuous. They reject complete-link clustering on criterion (a), and average-link clusterings on (b), and support single linkage as the only sorting strategy satisfying the conditions. To me their counter-examples show the limitations of the whole idea of classification and not merely of certain techniques. Continuity or stability is a property of the data and not an analytic property of the algorithm, unless the algorithm is expressed as a continuous transformation of the data (Jackson, 1970). Some—I am tempted to say most—data are just not classifiable. In one example Lange *et al.* (1965) found that even with single linkage "relatively small change in a percentage similarity [was amplified] in such a way as to impose large changes in the emerging pattern of cross-linkages" in a dendrogram.

The cophenetic correlation coefficient $\bar{D}1$ has been much used in practice to compare hierarchical representations. Many authors have reported that pair-group sorting L3 gives the highest $\bar{D}1$ of the sorting strategies tried (Boyce, 1969; Sneath, 1969). L3 has been often deliberately chosen in consequence (Mello and Buzas, 1968).

TABLE 3

Distortion measures

$\bar{D}1$	$\frac{\sum(s_{ij}-\bar{s}_{ij})(s^*_{ij}-\bar{s}^*_{ij})}{[\sum(s_{ij}-\bar{s}_{ij})^2 \sum(s^*_{ij}-\bar{s}^*_{ij})^2]^{\frac{1}{2}}}$	Sokal and Rohlf (1962) Kruskal and Carroll (1969)
$\bar{D}2$	$\frac{\sum d_{ij} d^*_{ij}}{[\sum d^2_{ij} \sum d^{*2}_{ij}]^{\frac{1}{2}}}$	Guttman (1968)
D3	$\sum[d^2_{ij}-d^{*2}_{ij}]$	Gower (1966, 1970)
D4	$\begin{cases} [\frac{1}{2}\sum \mid s_{ij}-s^*_{ij} \mid^{1/\mu}]^{\mu}, & 0<\mu\leqslant 1 \\ \max\limits_{i,j} \mid s_{ij}-s^*_{ij} \mid \end{cases}$	Jardine *et al.* (1967)
D5	$\sum w_{ij}[s_{ij}-s^*_{ij}]^2$	Hartigan (1967)
D5′	$\sum w_{ij}[d_{ij}-d^*_{ij}]^2$	Anderson (1971a)
D6	As D5′ with $w_{ij} = k$	Shepard (1962) Thompson and Woodbury (1970)
D7	As D5′ with $w_{ij} = 1/d_{ij}\sum d_{ij}$	Sammon (1969)
D8	$\sum[d^*_{ij}-f(d_{ij})]^2/\sum d_{ij}$ where $f(d_{ij})$ some "regression" function	Kruskal (1964) Kruskal and Carroll (1969)
D9	$\sum w_{ij}\, d^2_{ij}/d^{*2}_{ij}$ with $w_{ij} = \frac{1}{d^{*2}_{ij}[\sum(1/d^{*2}_{ij})]^2}$	Shepard and Carroll (1969)
D10	$N^{-1}\sum[d_{ij}/d^*_{ij}]^{2a}\frac{[N^{-1}\sum(d^*_{ij})^{2(a-b)}]^{b/(a-b)}}{[N^{-1}\sum(d_{ij})^{2(a-b)}]^{a/(a-b)}}$ with $a = \frac{1}{2}$, $b = 1$ or $a = b = \frac{1}{2}$	Kruskal and Carroll (1969)

For use see text: $\bar{D}$ signifies a coefficient that decreases with increasing distortion.
$\bar{D}1$, D4, D5: used initially to compare dendrograms or a similarity matrix with a dendrogram.
$\bar{D}2$, D3, D6–D10: used initially to compare geometrical representations in different number of dimensions.

Recently Farris (1969) has shown algebraically that these reports are correct: as measured by $\bar{D}1$, L3 is optimal. A proposal by Hope (1970) seems allied to $\bar{D}1$: he proposes to compare "matrices of dendrogram scores with one another by testing the regression of one set of scores on the other for significance, and by extracting the canonical roots and variates of the regression". No details are given, but the distributional and dependence problems seem immense.

Ordination

Another well-defined representation of a similarity structure is as a set of points in Euclidean $[p]$-space. If the initial data are in the form of an $(n \times p)$ data matrix or in

the form of a matrix of Euclidean inter-point distance matrix, representation S1 is immediate. If the initial characterization is as similarity structure S2 or S3 this can be represented in Euclidean $[n]$-space if the matrix $\mathbf{S}'(s'_{ij} = s_{ij} - \bar{s}_i - \bar{s}_j + \bar{s})$ is positive semi-definite (Gower, 1966, 1967a). This does not hold for the information-gain statistic (Williams *et al.*, 1969a). Characterization S4 can be represented as S3 by $s^*_{ij} = \frac{1}{2}(s_{ij} + s_{ji})$. Conditions for S5 to be representable as S2 are given by Beals and Krantz (1967). Ordination is the process of representing a characterization as points in Euclidean $[p^*]$-space, $p^* \ll p$.

Heuristic representations in $[p^*]$ where p^* is as small as possible were developed in ecology by Bray and Curtis (1957) and in psychology by several authors (see Shepard, 1962, for references). $\bar{\mathrm{D}}1$ has been used to assess such procedures (Swan and Dix, 1966). More statistically reputable were methods of factor analysis (Goodall, 1954b; Ihm, 1965; Sneath, 1968) and principal component analysis (advocated by Austin and Orloci, 1966). The geometric properties of these techniques were discussed by Orloci (1966), who pointed out the necessity for basing **S** on a centred similarity index (cf. Wishart, 1971), and more extensively by Gower (1966, 1967a). Gower uses the term "principal co-ordinate analysis" for his representation of **S** as n points in $[n]$ and their subsequent reduction to $[p^*]$ by the projection of the points P_i on to the hyperplane of the first p^* principal components. Such a representation minimizes D3. Examination of the individual "residuals" P_iQ_i, and of the angles P_iGQ_i (where Q_i is the projection of P_i and G the centroid of all points in both representations, cf. I4), may enable outlying points or clusters to be identified (Anderson, 1971a, b). Since outliers may seriously distort the principal axes, the analysis should be rerun with the outlying points removed. These analyses have the advantages that a representation in $[p^*]$ contains the representations in $[q]$ for all $q < p^*$ (Howard, 1969), and that new points can easily be incorporated (Gower, 1968).

Most similarity coefficients do not define an obviously optimal distance function. These functions which have been proposed (e.g. $1-s, \cos^{-1}s, -\log s$) are monotonically related. This idea of monotonicity forms the basis of Shepard's (1962) multidimensional scaling: the inter-entity distances in $[p^*]$ should be monotonically related to those in $[p]$. Kruskal (1964) introduced a measure of stress to represent the amount by which the representation in $[p^*]$ failed to meet this criterion. Shepard and Carroll (1966) later introduced the related idea of a continuous mapping from $[p^*]$ into $[p]$, developed further by Kruskal and Carroll (1969). These methods try to minimize distortions D6, D8, D9, D10 respectively. The higher p^* the smaller the distortion, so that some further criterion is needed for the minimum acceptable p^*.

These scaling techniques, being non-metric, have more flexibility than essentially linear techniques of principal components, principal co-ordinates or factor analysis. If the spatial forms of the characterized data are non-linear, these techniques will produce an acceptable representation in fewer dimensions than required by linear techniques. For examples of their success in extracting non-linear structural forms from multidimensional data, see Hodson *et al.* (1966), Shepard and Carroll (1966). A possible extension to polynomial principal components is introduced by Gnanadesikan and Wilk (1969). A geometrical method of comparing different mappings into $[p^*]$ is given by Gower (1970).

Although such ordination procedures are not classification methods, they may allow the data to be visually inspected and clusters found (Dupraw, 1964). If the optimal p^* is greater than 2 but not much greater, glyph techniques exist for picturing three- or four-dimensional data in two dimensions (Anderson, 1960; Hall *et al.*, 1968).

When the data have not been forced into clusters, the observer can assess better whether clusters exist. It has been argued that if only the ordering of the similarities is used, as in multidimensional scaling, the sudden change in similarities between two well-separated clusters will be disguised (Anderson, 1971a). However, it would be expected that where the similarities within each of two sets of points are all lower than those between the sets, multidimensional scaling would reveal two clusters. This situation may be informative about the optimal choice of weights w_{ij} in the measure of distortion: too much emphasis on correct representation of small distances may lead to a distorted overall picture. The view is often expressed (e.g. Williams and Dale, 1965) that ordination techniques do not reveal structure in the data. However, examples are reported by Boyce (1969) and Williams *et al.* (1969) in which principal components or co-ordinates provide clearer evidence, and understanding, of clustering than standard classification techniques.

Minimum variance clustering

With data characterized, or capable of being characterized, as n points in Euclidean $[p]$-space, we may seek a representation as n^* points in Euclidean $[p]$-space, these points being the type specimens or profiles of the clusters found. One set of optimality criteria which has been much used is that of functions of the between and within cluster covariance matrices, whose distributional properties are well known under standard normality assumptions. The centroids of the clusters act as profiles. Algorithms for centroid methods have been discussed earlier.

The most thorough discussion of the principles by which the criterion to be maximized should be chosen is given by Friedman and Rubin (1967). They avoid the problems of scaling and correlated variables by restricting themselves to criteria invariant under non-singular linear transformations of the variables in the original data matrix, $|\mathbf{W}+\mathbf{B}|/|\mathbf{W}|$ or $\mathrm{tr}\,[\mathbf{W}^{-1}\mathbf{B}]$, both of which can be expressed as functions of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. Such criteria do not permit the definition of inter-entity similarity and, by using a pooled within-cluster matrix $\mathbf{W}$, have an underlying assumption that all clusters have the same shape, even if not restricted to sphericity.

Local density

If entities are characterized as points in a metric space, a natural concept of clustering is that there should be parts of the space in which the points are very dense separated by parts of low density (Carmichael *et al.*, 1968). In one dimension, bimodality has long been taken as evidence of a mixture. No restriction should be placed on the shape of the dense centres. One approach is to partition the space and count points, but if the partitioning is done without reference to the points, as in the Cartet count (Cattell and Coulter, 1966), the scale of partitioning must be exceedingly coarse, since if each dimension is partitioned into z classes the n points are to be shared among z^p cells. More practicable is to consider each given point in turn and partition the space into a region "near it" and "far from it". A histogram of the frequency distribution of distances to the other $(n-1)$ points should have a mode near zero if the centre point is part of a cluster, and other modes may reveal further clusters (Johnson and Wall, 1969). But whether partition is made at a fixed distance r from the chosen point (Wishart, 1969a, c) or at the gth nearest neighbour (Harrison, 1968), a rotationally invariant local geometry is implied. A less severe restriction seems to be required by the algorithm for detecting unimodal fuzzy sets given by Gitman and Levine (1970).

Harrison proposes to test for existence of clusters by finding the probability, with binary data and assuming attributes independent, that, of the $(n-1)$ similarities with the central point, g should have similarity no less than the observed gth nearest. This proposal is closely related to that of Goodall (1964, 1966c). Most searches for local modes are in practice closely related to single-linkage sorting. Wishart (1969c) recommends single-linkage sorting of all points whose local density is above some threshold, the remaining "noise" points being allocated later.

Multimodality can be sought directly by considering mixtures of distributions, although, in view of the difficulty of expressing simply conditions for bimodality of a mixture of two univariate normal distributions (Eisenberger 1964; Behboodian 1970), the method may not be as productive as some authors have hoped (e.g. Ihm, 1965). The recent discussion by Day (1969) of properties of mixtures of multivariate normal distributions gives some hope for future development, renders obsolete arguments against seeking univariate bimodality that the wrong variable may be being studied (Hope, 1970), and provides a statistically sound test of clustering in this situation.

Choice of dimension

Common to both partitional clustering and ordination is the problem that even when optimal representation can be found with n^* clusters or p^* dimensions, choice of n^*, p^* has still to be made. No formal criteria have been laid down to incorporate aspects such as the loss in comprehension of an ordination with increasing p^*. What is common practice in many situations is to look for a discontinuity in slope in the graph of the minimized distortion against p^* or n^* (Thorndike, 1953; Kruskal, 1964; Ihm, 1965; Jancey, 1966; Friedman and Rubin, 1967).

Calinski and Harabasz (1971) choose the n^*-part partition of the minimum spanning tree which maximizes

$$C = \frac{\operatorname{tr}\mathbf{B}}{n^*-1} \Big/ \frac{\operatorname{tr}\mathbf{W}}{n-n^*}.$$

C increasing monotonically with n^* suggests no cluster structure; C decreasing monotonically with n^* suggests a hierarchical structure; C rising to a maximum at n^* suggests the presence of n^* clusters. Applied to Rao's (1952) data on Indian castes, this procedure suggests the absence of any cluster structure, with a little evidence of the existence of five groups identical with those postulated by Rao.

Reaching a decision on the basis of a discontinuity observed in the data is well known to be a hazardous procedure. If the set $\mathscr{E}$ is unique it is all that can be done. If $\mathscr{E}$ is in any sense a sample from a larger population, then the number of clusters or dimensions suggested by the first analysis can be treated as a hypothesis to be tested by new data.

5. MISCELLANEOUS FORMULATIONS

Graph theory

With suitable alterations in language most classification problems can be expressed in terms of graph theory. With similarity structure S10 the binary matrix is the adjacency matrix of a graph, and this structure can readily be obtained from a similarity matrix $\mathbf{S}$ by replacing s_{ij} by 1 or 0 according as it is greater or not greater than some threshold t (Bonner, 1964; Rose, 1964; Estabrook, 1966; Batty, 1969). The choice of t is often determined by the nature of the computer used (Parker-Rhodes and Jackson,

1969). Within this any connected subgraph corresponds to a single linkage cluster, a maximal complete subgraph to complete linkage. An algorithm for finding all maximal complete subgraphs is given by Cole and Wishart (1970).

Two procedures have been suggested for single-linkage clusters. As t increases connected subgraphs will become split, so that a breaking point t can be associated with every cluster. Any connected subgraph of r points must have at least $(r-1)$, at most $\frac{1}{2}r(r-1)$, lines. Estabrook (1966) and Wirth *et al.* (1966) define the connectedness of a cluster by

$$C = \frac{s-(r-1)}{\frac{1}{2}r(r-1)-(r-1)},$$

where s is the observed number of connections in the cluster. Its separation from other clusters is measured by its moat M, the amount by which t would have to be reduced for another point to be included in the cluster. Consideration of C and M should allow decisions to be taken on natural clusters, presumably by some discontinuity principle (cf. Rubin's stability, discussed above). Very similar criteria are given for overlapping clusters by Jardine and Sibson (1968b).

Rose (1964) determines those points and lines of any connected subgraph that are most likely to be cut-points and cut-sets by repeatedly taking random pairs of points, finding the shortest path between them, and building up randomization distributions of the number of times each point and each line is included. From this a significance threshold can be obtained, and the process repeated with new random choices until significant points and lines are found. These can then be removed and the remaining subgraph examined for connectedness.

To seek space–time clustering of diseases Pike and Smith (1968) following Knox (1964b) construct adjacency digraphs (since time at least is irreversible) for space and for time of associated events. The null hypothesis that the digraphs are independent can be tested by a randomization test of the points of one digraph on the other or by suitable approximation to this.

Of more interest because of the immense reduction in data achieved is the minimum spanning tree, formed by joining each point to its nearest neighbour, each group thus formed to its nearest group neighbour, and so on until one polygonal line, the shortest possible, links all n points (Florek *et al.*, 1951). The tree is unaffected by monotonic transformations of the distance function and gives rise to single-linkage clusters at any threshold t by breaking links of length greater than t (Gower and Ross, 1969). The adaption of this to centroid clustering by Calinski and Harabasz (1971) is discussed above. Plotting the minimum spanning tree (MST) on the results of another cluster or ordination analysis is recommended for displaying distortion in the representation (Gower and Ross, 1969; Thompson and Woodbury, 1970). The MST has also been used to handle sets of data too large for more conventional classification procedures (Ross, 1969).

Overlapping clusters

Most classification techniques lead to disjoint clusters. In biology this is usually regarded as a *sine qua non*, although minimum distortion of data is sometimes given priority (Dupraw, 1964). In language studies a word can have more than one meaning and hence can reasonably be a member of more than one cluster of words (Jones and Jackson, 1970). In such studies clustering is into a group set aside for further examination, and a group rejected: it is thus "intellectually odd to have complete symmetry

between a clump and its complement" (Jones and Needham, 1968). This is a philosophy which other workers, particularly in fields where co-presence is considered more significant than co-absence, might seriously consider.

Needham (1967) initially considered a symmetric cohesion function between a clump and its complement, $s^2_{c\bar{c}}/s_{cc}s_{\bar{c}\bar{c}}$, where $s_{AB} = \sum_{i\in A}\sum_{j\in B} s_{ij}$, s_{ij} taken as Jaccard's index I9 subjected to a threshold. This was found to yield small, well-separated clumps but the lack of weighting led to unsatisfactory results (Jones and Jackson, 1967). A modification

$$\frac{s_{c\bar{c}}}{s_{cc}}\left\{\frac{n_c(n_c-1)}{s_{cc}}-\frac{s_{cc}}{pn_c(n_c-1)}\right\}$$

by Jackson (Parker-Rhodes and Jackson, 1969) gives clumps which tend to be large and overlapping, although the parameter p allows the user some control. Algorithms to minimize these functions iterate to a local minimum by successive reallocation of single individuals from an initial random partition, or from an initial randomly chosen cluster centre (Jones and Jackson, 1967; Bonner 1964, 1966). All such algorithms waste time through the repeated finding of the same clump: no way to avoid this completely is known.

The clusters revealed by complete-linkage sorting can be represented as maximal complete subgraphs at any assigned threshold. Often these will overlap, and a disjoint set cannot uniquely be defined. These overlapping clusters are part of a connected subgraph found at the same threshold by single-linkage sorting. At a lower threshold both subgraphs will still be present, but possibly submerged in larger maximal complete subgraphs and longer connected subgraphs. A mathematical formulation of such a system in which not more than $(k-1)$ points are allowed in the overlap of two clusters was given by Jardine and Sibson (1968a). In Sibson (1970) axioms of stability, optimality, cluster-preservation and invariance under relabelling or any monotonic transformation of S are shown to lead uniquely to this formulation. The requirement of monotonicity ensures that only the logical operations of maximum and minimum, expressed in complete linkage and single linkage, can be included. The algorithm given by Jardine and Sibson (1968b) for implementing this method has been improved by Cole and Wishart (1970).

The effect of increasing k is to lessen the distortion between the data and the ultrametric of the hierarchic structure. When $k = n-1$ the distortion is zero, but understanding of the overlapping system of clusters also approaches zero. The effect of k can be assessed by measures of isolation and connectedness similar to these discussed earlier in this paper.

Second-order processes

A second-order concept of a cluster is obtained by requiring that

$$(s_{i1}, \ldots, s_{in}) \simeq (s_{j1}, \ldots, s_{jn})$$

if E_i, E_j are in the same cluster. Tryon (1939, 1958), clustering variables rather than entities, requires only that "in an ideal cluster the pattern of correlations of each variable is collinear with those of its mates" and adopts an "index of proportionality" $(\sum_k \rho_{ik}\rho_{jk})^2/\sum\rho^2_{ik}\sum\rho^2_{jk}$ (cf. D2) as his basic similarity index. Kendall (1963) adopts a more precise version of this approach. McQuitty (1967b) proposes that this process of treating column vectors of the matrix S as the co-ordinates of entities for which a similarity matrix S′ can be calculated should be iterated until $\mathbf{S}^{(r)}$ stabilizes. Bonner

(1964) had earlier made the same proposal, reducing each matrix to a binary one by subjecting it to some threshold. Fascinating sequences of matrices result, particularly if variables are rescaled each time, whose mathematical properties might repay investigation. In practice they seem to converge quickly but not always to non-overlapping clusters, but I know of no convergence theorem.

Derived structure analysis

Some early approaches to clusters regarded them as sets of entities within which variables or attributes are independent. The correlations between botanical species within an area implies that the area is heterogeneous (Goodall, 1953). Thus one should choose the species most highly correlated with the other species and partition the population of sample areas into two groups according to whether or not the area does or does not contain the key species. This was formalized by Williams and Lambert (1959, 1960) in association analysis. They measured the "information" contained by species u about the other species first by $\sum_{v\neq u}^{p}\chi^2_{uv}$ and later by $\sum_{v\neq u}^{p}\sqrt{(\chi^2/n)}$, summed over the $(p-1)$ 2×2 tables of presence and absence of species u and species v in the areas. The resulting monothetic hierarchical divisions were assessed by the maximum inter-species χ^2 found in each cluster. This measure does not decrease monotonically and does not have the distributional properties ascribed to it by Williams and Lambert. Moreover, "natural taxa do not in general result from such monothetic classifications" (Bailey, 1967; see also Sneath, 1965; Mandel 1969). However, the method was feasible and gave answers.

Lance and Williams (1968) have recently re-activated the method using the true information content instead of χ^2, which is an approximation to it (Macnaughton-Smith, 1965). But they repeat the earlier error of having a stopping rule based on assessing the maximum of a number of (correlated) χ^2 against the distribution of a single χ^2. The major criticism is that, since different parts of the hierarchy contain different numbers of variables, the stopping rule is not comparable between different parts of the same analysis. When users construct hierarchies "terminated at the conventional $P = 0{\cdot}05$ level" (Tracey, 1968) their results can be totally invalid. The same criticism applies to any technique for obtaining clusters by stopping an agglomerative or divisive hierarchical process.

An alternative scheme has been proposed by Crawford and Wishart (1967, 1968) for rapid assessment of a large-scale binary ecological matrix $\mathbf{X}_{iv}$. It is assumed that ecological groups are determined by species which occur frequently with high density, rather than those which are frequent but isolated, or occur infrequently in rich areas. For each species V_u its group element potential (G.E.P.) is calculated essentially as

$$W_u = \sum_{v=1}^{p}\sum_{i=1}^{p} x_{iv}\,x_{iu}.$$

The set element potential (S.E.P.) of an entity E_i is defined as $\sum_{u=1}^{p} W_u x_{iu}$ scaled to be less than 1 by division by the maximum S.E.P. A simple measure is defined of the interaction of S.E.P. and G.E.P. for each species and the population partitioned into entities which contain or do not contain the species maximizing this interaction. The classification obtained by terminating this monothetic hierarchical division at an arbitrary level can be improved by iterative reallocation of those entities which have low S.E.P. with respect to their cluster.

Latent structure analysis is a more direct approach to finding clusters within which variates are uncorrelated (Lazarsfeld and Henry, 1968). The co-occurrences of up to

k characters are required to allow estimation of allocation probabilities into k latent classes. The classes found are disjoint, but each individual is labelled with probabilities of belonging to the various classes, instead of being definitely assigned to one (Baker, 1962). Good (1965a) attacks the same problem via information theory. This formulation does not seem to have been used by biologists, but could be worth considering for such problems as disease diagnosis or land use.

Noda

Derived structure analyses make plain what can be forgotten in direct cluster analyses, that there is an underlying relationship between entities and variables. Hope (1969a, b) has argued strongly that this be borne in mind even when doing conventional analyses.

There have been few attempts to elicit directly from the data important combinations of entities and variables, possibly because it is very difficult to define precisely what is desired. "A nodum is an enumeration both of a set of points and of the set of axes in which the points constitute a galaxy" (Williams and Dale, 1965). The first, by Lambert and Williams (1962), clusters entities on the basis of inter-variable associations and variables on the basis of inter-entity associations allowing each process to be modified by the other. A simpler procedure is given by Tharu and Williams (1966) who consider a binary data matrix partitioned into m and $(n-m)$ entities, and q and $(p-q)$ variables. The four cells thus formed contain at most mq, $m(p-q)$, $(n-m)q$, $(n-m)(p-q)$ unit entities. The observed numbers in these four cells may be used to test the hypothesis that the proportion of 1's is the same in all four cells. The resulting χ^2_3 can be partitioned into terms representing the partitioning of individuals, the partitioning of variables and their interaction. Any of these terms, or their total, could become the criterion to be maximized.

Factor analysis models for binary data place entities and characters on a symmetric basis. Macnaughton-Smith (1965) considers various representations of the probability that entity E_i possesses character V_u as a function of two sets of parameters. He adopts $p_{iu} = \alpha_i \beta_u/(1+\alpha_i \beta_u)$ as a suitable model for the desired lack of interaction between characters and entities in a homogeneous cluster, and gives a first approach to an algorithm for finding clusters which display a good fit to this model.

Day (1970) generalizes this idea in a model,

$$p_{iu} = f\left[\sum_{r=1}^{k} a_{ur} b_{ir} + d_i\right],$$

where f can be taken as the logistic function. The parameter can be fitted by maximum likelihood successively for increasing values of k until a sufficiently good fit to the model is achieved. A unified geometrical representation of characters and entities is obtained. Entities are represented as points, variables as hyperplanes (or vice versa) in k-dimensional factor space. If $k \leqslant 3$ the inter-relations of entities and variables can be visually inspected.

6. CONCLUSION

Clustering uses much time and effort. We want to cluster only if clusters exist (Fleiss and Zubin, 1969) and it would be useful to have some test of this without going through the whole process of finding best clusters, which then turn out to be not good enough: the ability of procedures to find non-existent clusters is well

established (Forgy, 1965). A possible approach is provided by Hills (1969) for correlation coefficients ρ_{ij}. He applies the formal transformation $z = \frac{1}{2}\log[(1+\rho)/(1-\rho)]$ to normality and, ignoring non-independence, draws a half-normal plot of z. This can be repeated on within-cluster correlation matrices, and on a matrix of similarities between single representatives of each cluster. A gamma probability plot has been suggested (Gnanadesikan and Wilk, 1969) for bringing to light certain configurations of points. Suitable transformations and graphical representations of other similarity indices could be explored. Conservative methods should be encouraged. But their application to the lower branches of a hierarchy, assessing the properties of a group formed as a consequence of an earlier statistical test, poses problems of a kind which statisticians have, reasonably, tended to avoid.

Often the act of classification has a primary purpose. If so, that purpose should be taken into account. Special techniques have been proposed when the aim is to predict one of the variables (Macnaughton-Smith, 1963; Morgan and Sonquist, 1963) or to construct an identification key (Gower, 1969a). One possible development is to the study of temporal changes, either natural (Williams *et al.*, 1969a) or experimental (Tracey, 1968), in a complex set of multivariate entities. But if a specific question can be asked it is likely that standard multi-purpose classification techniques will give a poor answer. Easily observed variables may be little correlated with real structure: consider Mendeleev's periodic table (Muir, 1962). And the arguments about the ethics of weighting become irrelevant when it is realized that infinitely more weight is given to an observed variable compared to an unobserved one (Muir *et al.*, 1970).

Every point raised by Tukey (1954) in his general principles for statisticians has relevance for taximeters (i.e. practitioners of taximetrics; Johnson, 1968). Most users ignore three of his dicta: "Different ends require different means and different logical structures." "While techniques are important...knowing when to use them and why to use them is more important." "In the long run it does not pay a statistician to fool either himself or his clients." But how in practice does one tailor statistical methods to the real needs of the user, when the real need of the user is to be forced to sit and think?

One of Good's (1965b) reasons for classification was "for fun". Many people so regard it. In a brilliant, witty and utterly scathing discussion, Johnson (1968) pillories such "scientists" so accurately that I end with his words:

"Anyone who is prepared to learn quite a deal of matrix algebra, some classical mathematical statistics, some advanced geometry, a little set theory, perhaps a little information theory and graph theory, and some computer technique, and who has access to a good computer and enjoys mathematics...will probably find the development of new taximetric method much more rewarding, more up-to-date, more 'general', and hence more prestigious than merely classifying plants or animals or working out their phylogenies."

ACKNOWLEDGEMENTS

I am most grateful to the many friends and colleagues who have added titles to the list of references on which this paper is based.

REFERENCES

Starred references are principally survey papers with valuable bibliographies.

ANDERSON, A. J. B. (1971a). Numeric examination of multivariate soil samples. *Math. Geol.*, **3** (in press).

—— (1971b) Ordination methods in ecology. *J. Ecol.*, **59** (in press).

ANDERSON, E. (1960). A semigraphical method for the analysis of complex problems. *Technometrics*, **2**, 387–392.
AUSTIN, M. P. and ORLOCI, L. (1966). Geometric models in ecology. II. An evaluation of some ordination techniques. *J. Ecol.*, **54**, 217–227.
BAILEY, N. T. J. (1967). *The Mathematical Approach to Biology and Medicine.* New York: John Wiley.
BAKER, F. B. (1962). Information retrieval based on latent class analysis. *J.A.C.M.*, **9**, 512–521.
BALAKRISHNAN, V. and SANGHVI, L. D. (1968). Distance between populations on the basis of attribute data. *Biometrics*, **24**, 859–865.
*BALL, G. H. (1965). Data analysis in the social sciences: What about the details? In *Proc. Fall Joint Computer Conference, Stanford*, pp. 533–559. New York: Macmillan.
BALL, G. H. and HALL, D. J. (1967). A clustering technique for summarizing multivariate data. *Behaviour Sci.*, **12**, 153–155.
BANNISTER, P. (1968). An evaluation of some procedures used in simple ordinations. *J. Ecol.*, **56**, 27–34.
BARON, D. N. and FRASER, P. M. (1968). Medical applications of taxonomic methods. *Brit. Med. Bull.*, **24**, 236–240.
BATTY, C. D. (1969). The automatic generation of index languages. *J. Document.*, **25**, 142–151.
BEALE, E. M. L. (1969a). *Cluster Analysis.* London: Scientific Control Systems.
—— (1969b). Euclidean cluster analysis. *Bull. I.S.I.*, **43**, Book 2, 92–94.
BEALS, R. and KRANTZ, D. H. (1967). Metrics and geodesics induced by order relations. *Math. Zeit.*, **101**, 285–298.
BEERS, R. J. and LOCKHART, W. R. (1962). Experimental methods in computer taxonomy. *J. Gen. Microbiol.*, **28**, 633–640.
BEHBOODIAN, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics*, **12**, 131–139.
BOLSHEV, L. N. (1969). Cluster analysis. *Bull. I.S.I.*, **43**, Book 1, 411–425.
BONNER, R. E. (1964). On some clustering techniques. *IBM J. Res. Dev.*, **8**, 22–32.
—— (1966). Cluster analysis. *Ann. N.Y. Acad. Sci.*, **128**, 972–983.
BOULTON, D. M. and WALLACE, C. S. (1970). A program for numerical classification. *Comp. J.*, **13**, 63–69.
BOYCE, A. J. (1969). Mapping diversity: A comparative study of some numerical methods. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 1–30. New York: Academic Press.
BRAY, J. R. and CURTIS, J. T. (1957). An ordination of the upland forest communities of S. Wisconsin. *Ecol. Monogr.*, **27**, 325–349.
CAIN, A. J. and HARRISON, G. A. (1958). An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc. Lond.*, **131**, 85–98.
CALINSKI, T. (1969). On the applications of cluster analysis to experimental results. *Bull. I.S.I.*, **42**, Book 2, 101–103.
CALINSKI, T. and HARABASZ, J. (1971). A dendrite method for cluster analysis. *Biometrics*, (in press).
CARMICHAEL, J. W., GEORGE, J. A. and JULIUS, R. S. (1968). Finding natural clusters. *Syst. Zool.*, **17**, 144–150.
CATTELL, R. B. (1944). A note on correlation clusters and cluster search methods. *Psychometrika*, **9**, 169–184.
—— (1949). r_p and other coefficients of pattern similarity. *Psychometrika*, **14**, 279–298.
CATTELL, R. B. and COULTER, M. A. (1966). Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. *Brit. J. Math. Statist. Psychol.*, **19**, 237–269.
COCHRAN, W. G. and HOPKINS, C. E. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, **17**, 10–32.
COLE, A. J. and WISHART, D. (1970). An improved algorithm for the Jardine–Sibson method of generating overlapping clusters. *Comp. J.*, **13**, 156–163.
COLE, L. C. (1949). The measurement of interspecific association. *Ecology*, **30**, 411–424.
COLEMAN, J. S. and MACRAE, D. (1960). Electronic processing of sociometric data for groups up to 1000 in size. *Amer. Sociol. Rev.*, **25**, 722–726.
COLMAN, G. (1968). The application of computers to the classification of streptococci. *J. Gen. Microbiol.*, **50**, 149–158.

CONSTANTINESCU, P. (1966). The classification of a set of elements with respect to a set of properties. *Comp. J.*, **8**, 352–357.

CRAWFORD, R. M. M. and WISHART, D. (1967). A rapid multivariate method for the detection and classification of groups of ecologically related species. *J. Ecol.*, **55**, 505–524.

—— (1968). A rapid classification and ordination method and its application to vegetation mapping. *J. Ecol.*, **56**, 385–404.

CZEKANOWSKI, J. (1913). *Zarys metod statystycznych.* Warsaw.

DAGNELIE, P. (1960). Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bull. Serv. Carte Phyto.*, **5**, 7–71, 93–195.

*—— (1965). L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques: principes fondamentaux. *Biometrics*, **21**, 345–361.

*—— (1966). A propos des différentes méthodes de classification numerique. *Rev. Stat. App.*, **14**, 55–75.

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

—— (1970). A model for multivariate dichotomous data. *Proc. 7th Inter. Bioms. Conf.*

DE LA VEGA, W. F. (1967). Techniques de classification automatique utilisant un indice de ressemblance. *Rev. franç. sociol.*, **8**, 506–520.

DICE, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

DUPRAW, E. J. (1964). Non-Linnaean taxonomy. *Nature, Lond.*, **202**, 849–852.

EADES, D. C. (1965). The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. *Syst. Zool.*, **14**, 98–100.

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

EISENBERGER, I. (1964). Genesis of bimodal distributions. *Technometrics*, **6**, 357–363.

EL-GAZZAR, A., WATSON, L., WILLIAMS, W. T. and LANCE, G. N. (1968). The taxonomy of *Salvia*: a test of two radically different numerical methods. *J. Linn. Soc.* (*Bot.*), **60**, 237–250

ENGELMAN, L. and HARTIGAN, J. A. (1969). Percentage points of a test for clusters. *J. Am. Statist. Ass.*, **64**, 1647–1648.

ESTABROOK, G. F. (1966). A mathematical model in graph theory for biological classifications. *J. Theoret. Biol.*, **12**, 297–310.

FARRIS, J. S. (1969). On the cophenetic correlation coefficient. *Syst. Zool.*, **18**, 279–285.

FIELD, J. G. (1969). The use of the information statistic in the numerical classification of heterogeneous systems. *J. Ecol.*, **57**, 565–569.

FISHER, D. R. and ROHLF, F. J. (1969). Robustness of numerical taxonomic methods and errors in homology. *Syst. Zool.*, **18**, 33–36.

FISHER, W. D. (1958). On grouping for maximum homogeneity. *J. Am. Statist. Ass.*, **53**, 789–798.

—— (1969). *Clustering and Aggregation in Economics.* Baltimore: Johns Hopkins Press.

FLEISS, J. L. and ZUBIN, J. (1969). On the methods and theory of clustering. *Multivariate Behaviour. 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. *Colloq. Math.*, **2**, 282–285, 319.

FORGY, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21**, 768–769.

FORTIER, J. J. and SOLOMON, H. (1966). Clustering procedures. In *Proc. Symp. Multiv. Analysis, Dayton, Ohio* (P. R. Krishnaiah, ed.), pp. 493–506. New York: Academic Press.

FRIEDMAN, H. P. and RUBIN, J. (1967). On some invariant criteria for grouping data. *J. Am. Statist. Ass.*, **62**, 1159–1178.

GENGERELLI, J. A. (1963). A method for detecting subgroups in a population and specifying their membership. *J. Psychol.*, **5**, 457–468.

GITMAN, I. and LEVINE, M. D. (1970). An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique. *IEEE Trans. Comp.*, **C19**, 583–593.

GNANADESIKAN, R. and WILK, M. B. (1969). Data analytic methods in multivariate statistical analysis. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.), pp. 593–638. New York: Academic Press.

GOOD, I. J. (1965a). Speculations concerning the first ultra-intelligent machine. In *Advances in Computers* (F. L. Alt, ed.), Vol. 6, pp. 31–88. New York: Academic Press.

*—— (1965b). Categorization of classification. In *Mathematics and Computer Science in Medicine and Biology*, pp. 115–128. London: H.M.S.O.

GOODALL, D. W. (1953). Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Aust. J. Bot.*, **1**, 39–63.
—— (1954a). Vegetational classification and vegetational continua. *Angew. Pflanz. (Wien) Festsch. Aich.*, **1**, 168–182.
—— (1954b). Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Aust. J. Bot.*, **2**, 304–324.
—— (1964). A probabilistic similarity index. *Nature, Lond.*, **203**, 1098.
—— (1966a). Numerical taxonomy of bacteria—some published data re-examined. *J. Gen. Microbiol.*, **42**, 25–37.
—— (1966b). Hypothesis testing in classification. *Nature*, **211**, 329–330.
—— (1966c). A new similarity index based on probability. *Biometrics*, **22**, 882–907.
—— (1967). Distribution of the matching coefficient. *Biometrics*, **23**, 647–656.
—— (1968). Affinity between an individual and a cluster in numerical taxonomy. *Biometrie-Praximetrie*, **9**, 52–55.
GOODMAN, L. A. and KRUSKAL, W. H. (1959). Measures of association for cross classifications, II. Further discussion and references. *J. Am. Statist. Ass.*, **54**, 123–163.
GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.
—— (1967a). Multivariate analysis and multidimensional geometry. *The Statistician*, **17**, 13–25.
—— (1967b). A comparison of some methods of cluster analysis. *Biometrics*, **23**, 623–628.
—— (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, **55**, 582–585.
—— (1969a). The basis of numerical methods of classification. In *The Soil Ecosystem* (J. G. Sheals, ed.), pp. 19–30. London: Systematics Association.
*—— (1969b). A survey of numerical methods useful in taxonomy. *Acarologia*, **11**, 357–376.
—— (1970). Classification and geology. *Rev. I.S.I.*, **38**, 35–41.
GOWER, J. C. and ROSS, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.*, **18**, 54–64.
GRIGAL, D. F. and ARNEMAN, H. F. (1969). Numerical classification of some forested Minnesota soils. *Proc. Soil Sci. of Am.*, **33**, 433–438.
GUTTMAN, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, **33**, 469–506.
HALL, A. V. (1965). The peculiarity index, a new function for use in numerical taxonomy. *Nature, Lond.*, **206**, 952.
—— (1967). Studies in recently developed group-forming procedures in taxonomy and ecology. *J. South Afr. Bot.*, **33**, 85–96.
—— (1969a). Group forming and discrimination with homogeneity functions. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 53–67. New York: Academic Press.
—— (1969b). Avoiding informational distortion in automatic grouping programs. *Syst. Zool.*, **18**, 318–329.
HALL, D. J., BALL, G. H., WOLF, D. E. and EUSEBIO, J. W. (1968). Promenade—a system for on-line pattern recognition. In *The Future of Statistics*, (D. G. Watts, ed.), pp. 309–313. New York: Academic Press.
HARRISON, P. J. (1968). A method of cluster analysis and some applications. *Appl. Statist.*, **17**, 226–236.
HARTIGAN, J. A. (1967). Representation of similarity matrices by trees. *J. Am. Statist. Ass.*, **62**, 1140–1158.
HAYHOE, F. G. J., QUAGLINO, D. and DOLL, W. R. S. (1964). *The Cytology and Cytochemistry of Acute Leukaemias*. Spec. Rep. Ser. M.R.C. No. 304. London: H.M.S.O.
*HEYWOOD, V. H. and MCNEILL, J. (eds.) (1964). *Phenetic and Phylogenetic Classification*. London: Systematics Association.
HILLS, M. (1969). On looking at large correlation matrices. *Biometrika*, **56**, 249–254.
HODSON, F. R., SNEATH, P. H. A. and DORAN, J. E. (1966). Some experiments in the numerical analysis of archaeological data. *Biometrika*, **53**, 311–324.
HOPE, K. (1969a). The complete analysis of a data matrix. *Brit. J. Psychiat.*, **115**, 1069–1079.
—— (1969b). Complete analysis: a method of interpreting multivariate data. *J. Market Res. Soc.*, **11**, 267–284.
—— (1970). The complete analysis of a data matrix: application and interpretation. *Brit. J. Psychiat.*, **116**, 657–666.

HOWARD, N. (1969). Least squares classification and principal component analysis: a comparison. In *Quantitative Ecological Analysis in the Social Sciences* (M. Dogan and S. Rokkan, eds) Cambridge: M.I.T. Press.

HURLBERT, S. H. (1969). A coefficient of interspecific association. *Ecology*, **50**, 1–9.

IHM, P. (1965). Automatic classification in anthropology. In *The Use of Computers in Anthropology* (D. Hymes, ed.), pp. 357–76. The Hague: Mouton and Co.

JACCARD, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. vaud. sci. nat.*, **37**, 241–272.

JACKSON, D. M. (1969). Comparison of classifications. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 91–111. New York: Academic Press.

—— (1970). The stability of classifications of binary attribute data. *Class. Soc. Bull.*, **2**, 40–46.

JANCEY, R. C. (1966). Multidimensional group analysis. *Aust. J. Bot.*, **14**, 127–130.

JARDINE, C. J., JARDINE, N. and SIBSON, R. (1967). The structure and construction of taxonomic hierarchies. *Math. Biosci.*, **1**, 173–179.

JARDINE, N. (1967). The concept of homology. *Brit. J. Philos. Sci.*, **18**, 125–139.

—— (1970). Algorithms, methods and models in the simplification of complex data. *Comp. J.*, **13**, 116–117.

JARDINE, N. and SIBSON, R. (1968a). A model for taxonomy. *Math. Biosci.*, **2**, 465–482.

—— (1968b). The construction of hierarchic and non-hierarchic classifications. *Comp. J.*, **11**, 177–184.

*JOHNSON, L. A. S. (1968). Rainbow's end: the quest for an optimal taxonomy. *Proc. Linn. Soc. N.S.W.*, **93**, 8–45. (Reprinted in *Syst. Zool.*, **19**, 203–238.)

JOHNSON, R. L. and WALL, D. D. (1969). Cluster analysis of semantic differential data. *Educ. Psychol. Measur.*, **29**, 769–780.

JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

JONES, K. S. (1970). Some thoughts on classification for retrieval. *J. Document.*, **26**, 89–101.

JONES, K. S. and JACKSON, D. M. (1967). Current approaches to classification and clump-finding at the Cambridge Language Research Unit. *Comp. J.*, **10**, 29–37.

—— (1970). The use of automatically-obtained keyword classifications for information retrieval. *Inform. Storage*, **5**, 175–201.

JONES, K. S. and NEEDHAM, R. M. (1968). Automatic term classification and retrieval. *Inform. Storage*, **4**, 91–100.

KENDALL, D. G. (1963). A statistical approach to Flinders Petrie's sequence dating. *Bull. I.S.I.*, **40**, 657–680.

KENDALL, M. G. (1966). Discrimination and classification. In *Proc. Symp. Multiv. Analysis, Dayton, Ohio* (P. R. Krishnaiah, ed.), pp. 165–185. New York: Academic Press.

KENDRICK, W. B. and PROCTOR, J. R. (1964). Computer taxonomy in the Fungi Imperfecti. *Can. J. Bot.*, **42**, 65–88.

KENNARD, R. W. and STONE, L. A. (1969). Computer aided design of experiments. *Technometrics*, **11**, 137–148.

KING, B. F. (1967). Step-wise clustering procedures. *J. Am. Statist. Ass.*, **62**, 86–101.

KNOX, G. (1964). The detection of space-time interactions. *Appl. Statist.*, **13**, 25–29.

KRUSKAL, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

—— (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

KRUSKAL, J. B. and CARROLL, J. D. (1969). Geometrical models and badness-of-fit functions. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), Vol. II, pp. 639–671. New York: Academic Press.

KURCZYNSKI, T. W. (1970). Generalized distance and discrete variables. *Biometrics*, **26**, 525–534.

LAMBERT, J. M. and WILLIAMS, W. T. (1962). Multivariate methods in plant ecology: IV. Nodal analysis. *J. Ecol.*, **50**, 775–802.

—— (1966). Multivariate methods in plant ecology: VI. Comparison of information analysis and association analysis. *J. Ecol.*, **54**, 635–664.

LANCE, G. N. and WILLIAMS, W. T. (1966a). Computer programs for hierarchical polythetic classification. *Comp. J.*, **9**, 60–64.

—— (1966b). A generalized sorting strategy for computer classifications. *Nature*, **212**, 218.

—— (1967a). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comp. J.*, **9**, 373–380.

—— (1967b). A general theory of classificatory sorting strategies. II. Clustering systems. *Comp. J.*, **10**, 271–277.

LANCE, G. N. and WILLIAMS, W. T. (1968). Note on a new information-statistic classificatory program. *Comp. J.*, **11**, 195.
LANGE, R. T., STENHOUSE, N. S. and OFFLER, C. E. (1965). Experimental appraisal of certain procedures for the classification of data. *Aust. J. Bio. Sci.*, **18**, 1189–1205.
LAZARSFELD, P. L. and HENRY, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin Co.
LERMAN, I. C. (1969). On two criteria of classification. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 114–128. New York: Academic Press.
—— (1970). *Les Bases de la Classification automatique.* Paris: Gauthier-Villars.
LUCE, R. D. (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, **15**, 169–190.
*MCINTOSH, R. P. (1967). The continuum concept of vegetation. *Bot. Rev.*, **33**, 130–187.
MACNAUGHTON-SMITH, P. (1963). The classification of individuals by the possession of attributes associated with a criterion. *Biometrics*, **19**, 364–366.
—— (1965). *Some Statistical and Other Numerical Techniques for Classifying Individuals.* Home Office Research Unit Report No. 6. London: H.M.S.O.
MACNAUGHTON-SMITH, P., WILLIAMS, W. T., DALE, M. B. and MOCKETT, L. G. (1964). Dissimilarity analysis. *Nature*, **202**, 1034–1035.
MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th. Berkeley Symp.*, **1**, 281–297.
MCQUITTY, L. L. (1956). Agreement analysis: classifying persons by predominant patterns of responses. *Brit. J. Statist. Psychol.*, **9**, 5–16.
—— (1964). Capabilities and improvements of linkage analysis as a clustering method. *Educ. Psychol. Measur.*, **24**, 441–456.
—— (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educ. Psychol. Measur.*, **26**, 825–831.
—— (1967a). Expansion of similarity analysis by reciprocal pairs for discrete and continuous data. *Educ. Psychol. Measur.*, **27**, 253–255.
—— (1967b). A novel application of the coefficient of correlation in the isolation of both typal and dimensional constructs. *Educ. Psychol. Measur.*, **27**, 591–599.
MANDEL, M. (1969). New approaches to bacterial taxonomy: perspective and prospects. *Ann. Rev. Microbiol.*, **22**, 239–274.
*MAYNE, A. J. (1968). Some modern approaches to the classification of knowledge. *Class. Soc. Bull.*, **1**, No. 4, 12–17.
*MAYR, E. (1968). Theory of biological classifications. *Nature*, **220**, 545–548.
MELLO, J. F. and BUZAS, M. A. (1968). An application of cluster analysis as a method of determining biofacies. *J. Paleontol.*, **42**, 747–758.
MINKOFF, E. C. (1965). The effects on classification of slight alterations in numerical technique. *Syst. Zool.*, **14**, 196–213.
MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data. *J. Am. Statist. Ass.*, **58**, 415–434.
MUIR, J. W. (1962). The general principles of classification with reference to soils. *J. Soil. Sci.*, **13**, 22–30.
MUIR, J. W., HARDIE, H. G. M., INKSON, R. H. E. and ANDERSON, A. J. B. (1970). The classification of soil profiles by traditional and numerical methods. *Geoderma*, **4**, 81–90.
NAGY, G. (1969). Feature extraction on binary patterns. *IEEE Trans. Syst. Sci.*, **SSC5**, 273–278.
NEEDHAM, R. M. (1963). A method of using computers in information classification. In *Information Processing, 1962* (C. Popplewell, ed.), pp. 284–287. Amsterdam: North Holland.
—— (1965a). Computer methods for classification and grouping. In *The Use of Computers in Anthropology* (D. Hymes, ed.), pp. 345–356. The Hague: Mouton and Co.
—— (1965b). Automatic classification: models and problems. In *Mathematics and Computer Science in Medicine and Biology*, pp. 111–114. London: H.M.S.O.
—— (1967). Automatic classification in linguistics. *The Statistician*, **17**, 45–54.
NEEDHAM, R. M. and JONES, K. S. (1964). Keywords and clumps. *J. Document.*, **20**, 5–15.
ORLOCI, L. (1966). Geometric models in ecology: I. The theory and application of some ordination methods. *J. Ecol.*, **54**, 193–215.
—— (1967). An agglomerative method for classification of plant communities. *J. Ecol.*, **55**, 193–206.
—— (1969). Information theory models for hierarchic and non-hierarchic classifications. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 148–164. New York: Academic Press.

PARKER-RHODES, A. F. and JACKSON, D. M. (1969). Automatic classification in the ecology of the higher fungi. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 181–215. New York: Academic Press.

*PIELOU, E. C. (1969). *An Introduction to Mathematical Ecology*. New York: Wiley–Interscience.

PIKE, M. C. and SMITH, P. G. (1968). Disease clustering: a generalisation of Knox's approach to the detection of space-time interactions. *Biometrics*, **24**, 541–556.

RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley.

RAYNER, J. H. (1966). Classification of soils by numerical methods. *J. Soil Sci.*, **17**, 79–92.

RESCIGNO, A. and MACCACARO, G. A. (1961). The information content of biological classifications. In *Information Theory* (C. Cherry, ed.), pp. 437–446. London: Butterworth.

ROGERS, D. J. and TANIMOTO, T. (1960). A computer program for classifying plants. *Science*, **132**, 1115–1118.

ROGERS, D. J., FLEMING, H. and ESTABROOK, G. (1967). Use of computers in studies of taxonomy and evolution. In *Evolutionary Biology*, (T. Dobzhansky, M. K. Hecht, and W. C. Steere, eds), Vol. 1, pp. 169–196. New York: Appleton Century Crofts.

ROHLF, F. J. (1970). Adaptive hierarchical clustering schemes. *Syst. Zool.*, **19**, 58–82.

ROSE, M. J. (1964). Classification of a set of elements. *Comp. J.*, **7**, 208–211.

ROSS, G. J. S. (1969). Classification techniques for large sets of data. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 224–233. New York: Academic Press.

ROUX, M. (1969). An algorithm to construct a particular kind of taxonomy. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 234–240. New York: Academic Press.

RUBIN, J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. *J. Theor. Bio.*, **15**, 103–144.

SAMMON, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Computers*, **C18**, 401–409.

SAWREY, W. L., KELLER, L. and CONGER, J. J. (1960). An objective method of grouping profiles by distance functions and its relation to factor analysis. *Educ. Psychol. Measur.* **20**, 651–674.

SHEPARD, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, **27**, 125–139, 219–246.

SHEPARD, R. N. and CARROLL, J. D. (1966). Parametric representation of non-linear data structures. In *Proc. Symp. Multiv. Analysis, Dayton, Ohio* (P. R. Krishnaiah, ed.), pp. 561–592. New York: Academic Press.

SHEPHERD, M. J. and WILLMOTT, A. (1968). Cluster analysis on the Atlas computer. *Comp. J.*, **11**, 57–62.

SIBSON, R. (1970). A model for taxonomy: II. *Math. Biosci.*, **6**, 405–430.

SILVESTRI, L. and HILL, I. R. (1964). Some problems of the taxometric approach. In *Phenetic and Phylogenetic Classification* (V. H. Heywood and J. McNeill, eds), pp. 87–104. London: Systematics Association.

SNEATH, P. H. A. (1965). The application of numerical taxonomy to medical problems. In *Mathematics and Computer Science in Medicine and Biology*, pp. 81–91. London: H.M.S.O.

—— (1966). A comparison of different clustering methods as applied to randomly spaced points. *Class. Soc. Bull.*, **1**, No. 2, 2–7.

*—— (1967). Some statistical problems in numerical taxonomy. *The Statistician*, **17**, 1–12.

—— (1968). The future outline of bacterial classification. *Class. Soc. Bull.*, **1**, No. 4, 28–45.

*—— (1969a). Evaluation of clustering methods. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 257–271. New York: Academic Press.

—— (1969b). Recent trends in numerical taxonomy. *Taxon*, **18**, 14–20.

SOKAL, R. R. (1961). Distance as a measure of taxonomic similarity. *Syst. Zool.*, **10**, 70–79.

*—— (1965). Statistical methods in systematics. *Biol. Rev.*, **40**, 337–391.

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

SOKAL, R. R. and ROHLF, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, **11**, 33–40.

—— (1966). Random scanning of taxonomic characters. *Nature*, **210**, 461–462.

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

SWAN, J. M. A. and DIX, R. L. (1966). The phytosociological structure of upland forest at Candle Lake, Saskatchewan. *J. Ecol.*, **54**, 13–40.

THARU, J. and WILLIAMS, W. T. (1966). Concentration of entries in binary arrays. *Nature, Lond.*, **210**, 549.

THEIL, H. (1967). *Economics and Information Theory*. Chicago: Rand McNally.
THOMPSON, H. K. and WOODBURY, M. A. (1970). Clinical data representation in multidimensional space. *Computers and Biomedical Research*, **3**, 58–73.
THORNDIKE, R. L. (1953). Who belongs in the family? *Psychometrika*, **18**, 267–276.
TRACEY, J. G. (1968). Investigation of changes in pasture composition by some classificatory methods. *J. Appl. Ecol.*, **5**, 639–648.
TRYON, R. C. (1939). *Cluster Analysis: Correlation Profile and Orthometric Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor: Edward Brothers.
—— (1958). General dimensions of individual differences: cluster analysis versus multiple factor analysis. *Educ. Psychol. Measur.*, **18**, 477–495.
TUKEY, J. W. (1954). Unsolved problems of experimental statistics. *J. Am. Statist. Ass.*, **49**, 706–731.
WALLACE, C. S. and BOULTON, D. M. (1968). An information measure for classification. *Comp. J.*, **11**, 185–194.
*WALLACE, D. L. (1968). Clustering. In *International Encyclopaedia of the Social Sciences* (D. L. Sills, ed.), Vol. 2, pp. 519–524. New York: Macmillan.
WARD, J. H. (1963). Hierarchical grouping to optimise an objective function. *J. Am. Statist. Ass.*, **58**, 236–244.
WARD, J. H. and HOOK, M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educ. Psychol. Measur.*, **23**, 69–82.
WATSON, L., WILLIAMS, W. T. and LANCE, G. N. (1966). Angiosperm taxonomy: a comparative study of some novel numerical techniques. *J. Linn. Soc. (Bot.)*, **59**, 491–501.
WILLIAMS, W. T. (1969). The problem of attribute-weighting in numerical classification. *Taxon*, **18**, 369–374.
*WILLIAMS, W. T. and DALE, M. B. (1965). Fundamental problems in numerical taxonomy. In *Advances in Botanical Research* (R. D. Preston, ed.), Vol. 2, pp. 35–68. London: Academic Press.
WILLIAMS, W. T., DALE, M. B. and MACNAUGHTON-SMITH, P. (1964). An objective method of weighting in similarity analysis. *Nature*, **201**, 426.
WILLIAMS, W. T. and LAMBERT, J. M. (1959). Multivariate methods in plant ecology, I. Association analysis in plant communities. *J. Ecol.*, **47**, 83–101.
—— (1960). Multivariate methods in plant ecology, II. The use of an electronic digital computer for association analysis. *J. Ecol.*, **48**, 689–710.
WILLIAMS, W. T., LAMBERT, J. M. and LANCE, G. N. (1966). Multivariate methods in plant ecology, V. Similarity analyses and information analysis. *J. Ecol.*, **54**, 427–445.
WILLIAMS, W. T. and LANCE, G. N. (1965). Logic of computer-based intrinsic classifications. *Nature*, **207**, 159–161.
WILLIAMS, W. T., LANCE, G. N., WEBB, L. J., TRACEY, J. G. and CONNELL, J. H. (1969). Studies in the numerical analysis of complex rain-forest communities, III. The analysis of successional data. *J. Ecol.*, **57**, 515–536.
WIRTH, M., ESTABROOK, G. F. and ROGERS, D. J. (1966). A graph theory model for systematic biology. *Syst. Zool.*, **15**, 59–69.
WISHART, D. (1969a). Numerical classification method for deriving natural classes. *Nature*, **221**, 97–98.
—— (1969b). An algorithm for hierarchical classifications. *Biometrics*, **25**, 165–170.
—— (1969c). Mode analysis. In *Numerical Taxonomy* (A. J. Cole, ed.), pp. 282–308. New York: Academic Press.
—— (1971). A generalised approach to cluster analysis. Part of Ph.D. Thesis (1970), University of St. Andrews.

DISCUSSION ON DR CORMACK'S PAPER

Dr. M. HILLS (London School of Hygiene and Tropical Medicine): The topic which has been so ably reviewed this evening calls to mind, irresistibly, the once fashionable custom of telling fortunes from tea leaves. There is the same rather arbitrary choice of raw material, the same passionately argued differences in technique from one teller to another, and, above all, the same injunction to judge the success of the teller solely by whether he proves to be right. In the case of fortune tellers this usually led to good publicity when they were right and no publicity when they were wrong.