COENOSES 3 (1): 11-32, 1988
© 1988 by Edizioni LINT Trieste

KNOWING WHEN TO STOP: CLUSTER CONCEPT - CONCEPT CLUSTER

M.B. Dale*, Dept. Biologia, Università di Trieste, Via Valerio 30140, Trieste, I-34100 Italy
* Present Address: CSIRO, Division of Tropical Crops and Pastures, 306 Carmody Rd., St Lucia, 4067, Australia

Keywords: Classificatio, Cluster definition, Geometry, Graph, Significance, Tests, Concepts, Inference, Interpretation, Cladistics, Generation, System

Abstract: There have been many alternative ways in which clusters have been defined. Perhaps the most frequent choice has been the geometric concept of a cluster as a set of point ''close'' in some space, a concept related to notions of probability density functions and hence to the framework of mathematical statistics. However such a model is not everywhere suitable, and in this paper I shall also examine some of the alternatives, chosen from models which have been used in deciding the number of clusters present. The aim of this examination it twofold. Firstly to indicate what alternatives have in fact been suggested, for many of them are neither well-known nor widely applied. Secondly to try and explore the situations in which one definition might be more appropriate than another. Ultimately such a decision must rest with the analyst, or agent, for approaches to testing for existence of clusters, and for determining the number of clusters, are closely related to the nature of the clusters being sought.

Introduction

''The human mind has an invincible tendency to reduce the diverse to the identical'' (Huxley 1937), and one of the commoner methods of accomplishing such reduction is clustering. Unfortunately, clustering has most often been defined only loosely as *''a process of generating classes whose members are alike, while the classes themselves are differentiated one from another''*, although even this definition is not acceptable universally as we shall see. And such a definition does not really help us to choose between the various methods of clustering which are available. As Dubes and Jain (1976, 1979, 1980) point out in their reviews of validation methods, we need to know something more. Which methods are simply algorithmic variants? Which are useful in this or that situation? What in fact does this or that method seek to do? How good are the clusters identified? The loose definition is insufficient to aid us in answering such questions.

In fact, if greater precision of definition is desired, then there is much variety and conflict in the definition of clustering. So many that Gasking (1960), somewhat despairingly, has concluded that there is no single universally acceptable one. In fact he concluded that, in some cases at least, the notion of a cluster is really *''a bureaucratic decision unrelated to the objects which are clustered''!* His example of this is a convoy of ships, whose existence is independent of the ships composing it and of their spatial disposition on the sea. Indeed it can exist, bureaucratically, without any ships at all. Pattern for an agent indeed!

In such a situation we can only investigate whether conditions can be found which determine when some particular definition is likely to be appropriate. For example, many of the more statistical approaches, which seek to base clustering in a firm probabilistic frame, adopt a particular view of cluster as *modal mixtures*. (One hopes that this is not solely because this definition has some tractibility mathematically.) What is certainly *not* clear is that such a definition is relevant to all the problems which clustering is supposed to address. Certainly in some cases clusters are required and the problem is to find optimal partitions; a one dimensional example is the work of Perillo and Marone (1980 a, b). Again, the controversy in taxonomy between pheneticists and cladists can be regarded as essentially an argument over the applicability of the modal view of clustering, with pheneticists largely adopting the model in order to initially *distinguish* species as clusters of individual specimens, while cladists adopt another model based on presumed evolutionary generating processes in order to represent the sibling relationships *between* already defined species. It is clear that, in this case, there are at least two classes of problem which are really being addressed, so that two different clustering methods might well be appropriate. This would mean that finally the user still has to decide what it is he or she wishes to accomplish. Yet rational choice is only possible if the alternatives have been clarified. What, for example, is the problem addressed by Chiu and Wong (1986) when clustering to organise knowledge bases for expert systems?

Even if we can accept a geometric model with the objects being clustered represented as points in some dimensional framework, the modal mixture model itself may not be relevant. Clusters may not consist of points but of lines as in the work of Bezdek et al. (1981) and Bezdek and Anderson (1985) or even, in part, empty space! Thus Langridge (1971) has suggested that people naturally associate a certain surrounding area with a letter, and will willingly place points "inside" letters which contain no closed curves, for example *L* or *V*, which is clearly a topological impossibility! The letters can be regarded, if digitised, as a cluster of points and they are "capturing" some associated empty space. Whether this should be called clustering is perhaps a matter of dispute, yet the concept of a distinguishable region in space is clearly apparent even if it is a little diffuse in density at the edges. So it would seem that some explication, at least, is needed.

There is a considerable problem in attacking this question, for there is a very large literature to survey, and the objectives of any particular method of clustering are often not clearly stated. It is slightly more manageable to examine proposals for estimating *how many* clusters are actually present, if any. In such procedures we are naturally deciding that such clusters as we have are, or are not, adequate. If we decide to distinguish a(nother) cluster, then we must have decided that something was previously inadequately represented. Note, too, that I am not concerned with the properties of particular algorithmic clustering procedures, for example with an axiomatisation of agglomerative clustering methods (Lukasovà 1979, Gordesch and Sint 1974), but with what *kind* of cluster is being sought. This may be intimately bound up with the algorithm although often the same general search methods are appropriate to several different models. Nor is the complexity of the calculation at issue, for in most cases suboptimal solutions are the only reasonably attainable ones. In some special cases integer or dynamic programming methods implemented on very fast computers might provide optimal solutions at a tolerable cost, or special purpose parallel processors might become available but these do not alter the objective of the search.

Pragmatically, various rules of thumb can be used to decide the number of clusters which should be obtained. For n entities one rule in the TAXON package (Ross 1979) is:

$$K = (\sqrt{n} + Log_2(n))/2$$

which seems to slightly overestimate the number of clusters most users find acceptable. While not denying the usefulness of such heuristics, a more formal mechanism would seem to be desirable. Given the variety of objectives in clustering, no single technique seems likely to meet all situations, so we shall have to be content with alternatives. This is no real problem if we understand the basis on which the rules are predicated. There have in fact been a variety of empirical investigations of the sampling properties of various measures of "quality of clustering", including Gower and Banfield (1978), Sneath (1980), Milligan and Mahajan (1980) and Milligan (1981) though not all of these provide direct tests for the number of clusters.

In this paper I have therefore examined a variety of alternative proposals for formally determining whether to cluster, what form the clustering should take and how many clusters to accept as valid. From these I have tried to determine some organisation of the various proposals and of the models they imply. This organisation is not intended to be composed of totally disjunct classes, however. In some cases the proposals involve several different criteria, *i.e.*, are themselves composite. An obvious example of this is given by Yolkina and Zagorniko (1978). In other cases the criteria themselves are correlated; for example notions of probability can be usefully employed in various ways, but I do not have a class of 'stochastic' models *per se*; the probabilities are derived in a manner finally dependent on the model so that such a class distinction would add nothing. There are quite acceptable methods existing for deciding the number of clusters which do not use probability notions at all, and much debate as to whether such notions are necessary or desirable. The classes form only a framework for discussion of the implied models.

Altogether I have distinguished 10 such major classes with an eleventh raghag to contain some rather isolated approaches. Containment in this eleventh class should not, however, be regarded with disfavour; arguments *ad populum* have no place here. The classes distinguished are as discussed under the following headings:

1. "The true partition revealed".
2. Dialectics.
3. The Naked Ideal.
4. On talking to angels.
5. A la mode.
6. Introvert/extrovert.
7. The curate's egg.
8. It's who you know.
9. Meaningful relationships.
10. Generation and Geneology.
11. Gallimaufry.

I have also attempted to provide a broad coverage, though not an exhaustive one, since many variations on the themes appear in a widely scattered literature. The various classes seem to form a counterpoint, several considering the same kinds of problems from their various viewpoints, although it must be confessed that, in the musical analogy, the timing has been somewhat aleatoric, or better intuitive, since it is at the discretion of the players. I have also tended to be somewhat profligate with reference material, largely because the alternatives are widely scattered through a wide range of disciplines and some attempt to link various approaches seemed desirable.

"The true partition revealed"

One method commonly proposed for establishing the efficacy of some classification procedure is to generate some data with "known" structure and seek to determine if the classification can find this. Jackson (1970) has argued that an effective procedure for classification

1. will give a unique results,
2. will be unaffected by any pre-ordering of the entities being classified,
3. will be invariant to scale changes,
4. will be resistant to small changes, perturbations, of the data.

While these suggestions look innocuous enough, they are not quite so self-evident as might seem at first glance; their acceptance can have some unexpected consequences. First we may actually wish to examine several results, if they are all more or less equivalently acceptable or indeed ambiguous. Additionally we may wish to examine differences between algorithms, as do Jain, Indrayan, and Goel (1986) and D'Andrade (1978). Most algorithms, even when they have all the data available for examination, in fact provide suboptimal solutions if only because the search problems themselves are computationally hard. Different methods of searching for solutions can well lead to different results providing different views of the implicit data structure. Jackson's first condition has some disadvantages.

Second, although suboptimal solutions are the usual result, it is only in sequential classification methods, such as those of Uttley (1970), López de Màntaras and Aguilar-Martin (1985) or Davis (1985), that the order of presentation of the entities being classified is likely to have any marked biasing effect. Indeed it is known that some induction problems can only be solved if the data are presented in a specific order; this would mean both that any method which attempts to identify clusters in data presented sequentially may need to prescribe the order of presentation of the data and that methods which are order independent need not necessarily be preferable to sequential methods. In practice order does matter to human classifiers; a fruit soup may closely resemble a dessert, from which it is distinguished only by its (partial) order of presentation. Jackson's second condition is therefore not universally applicable.

Third, it is not in the least obvious that in all classifications we should wish to ignore scale changes. This is equivalent to saying that size is never important, under any circumstances. It is true that very often we choose to avoid problems of scale by appropriate normalisation, but this is rather different from claiming that we should never regard size as of significant interest. Size is often used as a surrogate for age, for example, and we might not choose to disregard it then. What Jackson seems to mean is that arbitrary changes in units of measure, miles to millimeters for example, should not modify the classification, and in those circumstances his suggestion is almost always acceptable. Even in this case the nature of the units might be important. The use of "feet" instead of "metres" might be acceptable as a descriptor which distinguishes, say, nationality of the author of the data but that is plainly a different situation.

The fourth criterion is an argument for equal weighting of attributes, and this is known to be unacceptable. Watanabe's (1969) "ugly duckling" theorem shows that there must be selection of attributes if patterns are to have any meaning at all! Furthermore even when an acceptable subset of attributes has been selected from the infinity available, there have been strong arguments for differentially weighting attributes. Indeed one might say that the real argument is about how this should be done, not whether it should be done. In this case a small change in an important attribute might correctly cause a large change in the classification, while large changes in unimportant ones have little effect. And there is a symmetry here in that the items being classified need not all be uniquely assignable to a single class. Wong and Liu (1975), for example, attempt to identify the degree of typicality of items analogous to the importance of attributes. We shall return to this notion later when discussing fuzziness in classification.

Jackson (1972) later continued his studies by examining the effects of perturbation of the data, a subject earlier addressed by Minkoff (1965). In effect, with binary data, some '1' values were changed to '0', and some '0' to '1', and the classification repeated on the so-modified data. But this leaves us with a problem of comparing the two results. There certainly exist methods of comparing partitions and trees; for example Ecob (1978), Milligan and Isaac (1980), Milligan, Soon and Sokol (1983), Verhelst, Koppen and Van Essen (1985), Hubert and Arabie (1985) and, somewhat more esoterically perhaps, Michaud (1983) and Roger (1978). The problem is whether such comparison is always interpretable and acceptable; Edelbrock (1979) has shown that it may not be so, which somewhat counter-intuitive conclusion is perhaps best illustrated with an example.

Let us consider a set of data consisting of 10 a's and 10 b's, each described by some set of binary characters. Our ideal classification will presumably separate these into two classes, and it is these classes we expect our classification of perturbed data to recover. Let us now perturb a single one of the a's to be an A. Our classification will likely identify this as an outlier first, and will therefore not recover the partition into a and b that we expect. However if we accept a further division, thus forming 3 groups not 2, then we shall again recover the classes we desire, together with information on the outlier. We have more information than we expected. Is this desirable or not? And should I fault the classification method for finding an outlier, but requiring three groups to match the original, two group, data? Or should I fault the example because it confuses the classification of the glyphs with the classification of the letters which they represent, irrespective of font.

The problem of course is that our perturbed data no longer possess the properties which we expect to dominate the class formation, so that our 'true' structure is not the only structure present. Whether such a possibility is relevant needs to be considered before the results of such 'comparisons with the true structure' are uncritically accepted. With some models, and notably the normal mixture models which have dominated much statistical thinking, it may be acceptable to ignore the outlier; it is regarded as 'unclassified'. This would mean that such studies as Basford and McLachlan (1985a) on the correctness of allocation would be of interest for establishing the effectiveness of classifications. It would seem that acceptability here has something to do with a notion of continuity and 'smooth' error structure. If categorical changes are possible then the notion of comparison with a 'true' result is probably unhelpful. Jackson (1972) has noted that the conditions of his specific test, with equiprobable and symmetric changes of state, are unlikely to be present in practice. He argues that he is testing self-consistency of the classification rather than any reality of clusters, a statement whose humility could be well copied by some other workers.

Another appeal to perturbation was made by Jardine and Sibson (1968) in trying to establish the preeminence of single linkage clustering. They perturbed a single element of a dissimilarity matrix. Now this is difficult to accomplish if the dissimilarities are themselves calculated from other descriptive data. How can a change in these underlying data cause a change in a single dissimilarity coefficient. And if more than one change results, perhaps the whole cluster structure should change catastrophically, rather than the continuous deformation envisaged by Jardine and Sibson. Single changes may be possible if the dissimilarities are directly observed but even then it may not be desirable to insist on a 'smooth' change in a classification. Jardine and Sibson's approach does reflect their interest in analysis of similarities between populations, where perhaps such deformations could also be found, but this is then a study in discrimination. In any case in the B_k clustering methods they seem less concerned to form classes than to represent the dissimilarities better by using larger k values, thus moving away from single linkage towards complete linkage. This is not the only path between the two extremes; Hubert and Baker (1977), d'Andrade (1978) and even Lance and Williams (1967) provide other routes.

While Jackson provides an archetypal approach, there are other possibilities to be considered. Feoli and Lagonegro (1984) discuss sampling effects on the performance of their intersection analysis. In part the concern is with the representativeness of the sampling. If certain environments are in fact more common than others, then we may well take more samples from them than from other, rare, environments. This would result in a clustering of the points, and results from sampling in a spatial frame of reference while interpreting in an environmental one. Resistance to this kind of problem is difficult to conceive, since it requires knowledge of the results in order to stratify appropriately. The proposals of Breiman *et al.* (1984) for significance testing in predictive clustering are also relevant here, although these authors use other methods to determine the number of classes. The significance testing is of the entire hierarchical structure.

Some workers have proposed examining the results of cluster analysis where the *real* clusters are presumed known. For example, Golden and Meehl (1980) test 6 methods to determine if sex differentiation is recovered. This of course immediately assumes that sex differentiation is the dominant structure, so dominant that it alone is worth recovering from the data. This would seem to be a somewhat large assumption! *The data may contain several competing structures, any or all of which might be of interest.* A method which performed well in recovering structure in isolation is not guaranteed to perform equally well in the presence of multiple structures though probably we will have some increased confidence in its ability to perform. In practice we often appreciate having multiple possibilities. Certainly this would apply to the proposals of Lee, Slagle and Mong (1976) for estimating missing values, to O'Callaghan's (1976) search for perceptual structure in images or Scher, Shneier and Rosenfeld (1982) seeking to identify collinear line segments.

Dialectics

Continuing with the spatial framework for sampling and the perturbatory problems of Feoli and Lagonegro (1984), here we attempt to establish the probity of our classes using other evidence which was not used in the classification process developing the groups. This is the very common approach of interpreting one organisation of the data in terms of another, of which a typical example would be the study of Little and Ross (1985). This is often a very profitable exercise, for the additional variables permit us to develop a semantics to combine with the syntax of the classes. Indeed, this would seem to be the only way in which we can finally accept that our clusters are real; we are able to *explain the meaning* of the observed clusters in the context of other information. Since we can have several sets of additional information, each and all of which might contribute to our understanding the importance of the agent classifying is apparent. Only that agent can finally determine which of the alternative interpretations is really relevant to the problem being addressed.

This interpretive role is related to the problem of *simultaneous* clustering in two sets of variables addressed for example by Mantel (1957), Klauber (1975) or Siemiatychi (1978); in both cases the clusters are derived from one set of attributes and imposed on the other. It must be distinguished both from the constrained clustering envisaged by Dale and Walker (1970) and later extended by Margules, Faith and Belbin (1985), and others, to permit multiple constraints to be simultaneously applied and from the extrinsic classification methods of Williams and Dale (1965). The studies of Rachman and Koz'yakov (1986) are also relevant here.

In the former methods, similarity in some set of attributes, for example spatial position, is a prerequisite and treated as binary in nature. Items without such similarity are simply *not* considered for potential membership of a cluster. It is possible to relax such constraints somewhat by simply making it more difficult for an item to join a cluster unless it has the required property. This is using a continuous rather than a categoric constraint, and is related to the notion of fuzzy sets. Thus we might ask not that items in a cluster are all adjacent in space but only that the cluster of items show spatial aggregation (or disaggregation *c.f.* Dale *et al.* 1984) as measured by some suitable index. But since in all these cases we are employing both sets of information we cannot regard the attainment of the required conditions as a *test* of cluster validity. The space-time clustering statistics of Mantel and others noted above does provide a test, albeit one whose small sample properties are largely unknown, and which has a very slow convergence to asymptotic limits.

In the latter case we would adopt an explicitly extrinsic clustering method because we do wish to predict values for some set of variables from another set (*c.f.* Williams and Dale 1965, Macnaughton-Smith 1965, Sonquist, Baker and Morgan 1973, Breiman *et al.* 1985, Klopman and Macina 1985). In these cases we can perhaps use special tests, such as jack-knife statistics, bootstrapping or the other tests associated with AID (see *e.g.* Scott and Knott 1976), to establish the efficacy of our classes probabilistically. It is interesting, however, that Breiman *et al.* use another, nonprobabilistic, measure to determine the number of classes, a balance between predictability and complexity which we shall meet again later. Furthermore, in an approach to prediction such as is adopted in the GUHA algorithm (Hájek and Havránek 1978) which separates possibly overlapping clusters one at a time in a clumping process, multivalued inductive logics are used without a probabilistic framework, although there are obvious relationships to particular formulae in probability theory. But first we have to determine what model we expect the other variables to fit so that we can see if we are identifying something interesting; human interpretors are adept at unwarranted reification and will easily interpret erroneous or random results! Thus, far from avoiding the problem of the nature of our clusters, we have just transferred, and possibly doubled it.

The usual model adopted is the standard ANOVA model with all variances equal and perhaps some differences in mean, but there can be problems here, especially in an ecological context. If we proceed in a straightforward manner, with perhaps a *nonparametric* analysis as a genuflection to the distributional problems, we shall identify variables which are *necessary* discriminators. These are variables which are required to identify the group but do not of themselves uniquely identify it. If instead we follow Ratkowsky and Lance (1978) in normalising the contributions, in their method just so that mixed attribute types can be compared, we find *sufficient* variables predominating. These are variables capable of distinguishing the group uniquely *if they are present* but in many cases may not observed. Ideally we should like *necessary and sufficient* characteristics, but these are only found in comparatively few cases. It seems in most problems we will have to choose one or the other.

The situation can be further complicated if there are missing values in the data, for then even within a single class, the subgroups which have values recorded may change with each attribute and the value of the attribute as a distinguishing feature is confounded with the missing information. We must also add the problems of inapplicability where not only is the variable not recorded, it is logically impossible for it ever to be recorded. You will find no feathers on a fish however hard you look! Yet such characteristics are in many cases the most desirable.

In some cases we can employ specialised tests, such as examining spatial distributions for aggregation and disaggregation. Krishna-Iyer's (1949) test was used in this manner by Dale *et al.* (1985). In others we may desire not the total variance but some residual remaining after fitting some further model, *e.g.* a linear regression, to be minimised. Both Sonquist, Baker and Morgan's (1973) AID, and Macnaughton-Smith's (1965) multiple predictive analysis implement various forms of the latter approach and both could be generalised further.

So far we have considered evaluation using several sets of attributes, but one set of objects. We can also attempt to evaluate our classification with reference to other objects. One method for doing this is given by Sandland and Young (1979). The idea here is that replicate samples should most often end up in the same groups; if your classification separates replicates into different classes then you have probably gone too far with your class formation. The Sandland-Young test is not perfect, since it requires *all* replicates to remain together in a single group, whereas it would be nice to give some importance to situations where most of the samples remain together. The effect of this is to produce too few groups in problems with large numbers of replicates. But the important questions lie in the definition of replicates; Sandland and Young used paired trawls sampling spatially close water columns.

Another method using objects is due to Smith and Jain (1984) who adapted the Friedman-Rafsky (1979) test based on the minimal spanning tree. Basically they argue that if the data are clustered then, on the minimal spanning tree, most objects should be connected to other objects with the *same* class label, whereas if clustering is not present then the class labels of objects adjacent on the tree should be *randomly* distributed. Smith and Jain then suggest that given *n* objects which it is desired to cluster, a further *n* should be generated from the same attribute space, at random, and the minimal spanning tree test applied to test if the original *n* objects show clustering with respect to this additional random set. The generation of a true random set would be difficult so Smith and Jain use an approximate method for determining the attribute space based on convex hulls; we might expect that at least some of the random points could lie outside the convex hull. A similar approach is developed by Panayirci and Dubes (1983) based on the Cox-Lewis statistic instead of the Friedman-Rafsky. It would also be possible, of course, to use any replication present to determine the several sets, thus avoiding the problems of having to generate a sample from the same attribute universe. Replicates could be expected to remain connected and have the same class label.

What do these approaches tell us about the models implied? The Sandland-Young (1979) test does seem to be relatively independent of any proposed model if only because we must choose our replicates before we cluster at all. The other object-based approach still seems to regard clusters geometrically, though it is true that objects which are close need not be directly connected on the three.

The Naked Ideal

In this approach, the user specifies an ideal cluster structure, and leaves the program to separate the clusters, estimating for each any parameters which the user may have incorporated in the model. By far the commonest explicit model is one of multivariate normal mixtures, as in the work of Day (1969 a, b), and this is often implicitly used as well, but there are a large number of other models of the ideal group. For example we might ask only that each entity be placed so that most of the entities near it, in whatever sense the agent cares to specify, are placed in the same class. Hsu, Walker and Ogren (1986) provide an example of choosing with some range of distributions when these are suitably parametrised, which would seem a very desirable approach although as yet there is little experience with the computational problems. The approach is commonly associated with relaxation algorithms (reallocation procedures) and with some divisive methods; it is less easily recognised with transposed structuring methods such as that of Williams and Bunt (1980).

In general, similarity is here defined for a set of entities, being the fit that they have to the desired model; this may use pairwise similarities but is not restricted to them alone. For example Kashyap and Oommen (1983) illustrate how similarities for sets of strings can be related to pairwise similarities for members of the set while Sneath (1966) and Gotoh (1986) use triples of points to define a measure of local similarity. Ozawa (1983) combines pairwise measures with measures associated with single points, a feature also of Wong and Liu's (1975) methods for defining typicality. But what is 'ideal' is clearly problem specific; Pirktl (1983) clustering computer programs will not in general have the same requirements as Klopman and Macina (1985) in clustering chemical compounds. Frid (1970) minimising a function over a tree, Ganasalingam and McLachlan (1979, 1980) using mixture methods and maximum likelihood estimation.

In fact a very large number of alternatives have been proposed as the ideal structure, though few have found general acceptance. Minimal variance, (see *e.g.* Sen Gupta 1982/83, Česka and Roemer 1971) and various related models dominate almost entirely especially in association with agglomerative algorithms, but multiplicative models might also be used as in the work of Gilbert and Wells (1966). Cliff *et al.* (1986) produce overlapping clusters. Sneath's (1966) local linearity for curve seeking is another example, with an algorithm for clumping points around the linear segments, using the triples-similarity measure noted above. The simplest examples seem to those of P. Hartigan (1985, see also Engelman and Hartigan 1969 and J. Hartigan 1985) who proposed a dip test for unimodality for a single variable, and the proposals of Baroni-Urbani and Buser (1976), Baroni-Urbani (1980) and Rousseau (1978) for various simple models of binary data. Lefkovitch (1975, 1976, 1978, 1980, 1982) has proposed fairly elaborate methods of searching for fit to idealised cluster models which might be more widely applied now that sufficiently powerful computational facilities are becoming available.

Unfortunately in practice things are decidedly more complicated, for the definition of an ideal cluster is not trivial. Often it is necessary to deal with structured data, correlation of attributes, arrays of values treated as a single attribute, logical dependency (where the existence of a record for one attribute depends on the value of another), quasi- and pseudo-metricity (Lance 1970), strings, polygons, trees, and so on. It is difficult enough to measure pairwise similarity with such a plethora of types, though much can be done through the notion of minimal mutation distances. Extension to more than pairwise collections seems extremely difficult, as can be seem from the work of Gotoh (1986), and others referenced therein, on associating 3, or more, strings.

It seems likely that minimal variance is not wanted very often in practice especially if only differences in means are acceptable, with assumed similarity of shape of clusters. Macnaughton-Smith (1965) provides a number of measures based on Information theory which are applicable to state data, although he associates these with significance test in a limited manner. More general models would include fuzzy assignment of entities to clusters, fuzzy and crisp covariance to allow for shape diversity as in Diday and Govaert (1974), and more complex diversity measures such as the two-parameter models (Dale and Anderson 1973). But the ideal is very much agent-oriented. Eigen, Fromm and Northouse (1974) had methods based on identifying modes in histograms of every attribute with cross-classification to obtain the final clusters. Basford and McLachlan (1985b) have 3-way mixtures, Dale and Webb (1975) inosculate search and a two-parameter model interacting, Dale and Walker (1970) introduced constrained classification, while Margules, Faith and Belbin (1985) showed how more generalised constraints could be uniformly introduced into an agglomerative algorithm. Le Quesne (1974) identifies the properties necessary for an acceptable cladistic model (see also Felsenstein 1983 for a discussion of statistical issues in a cladistic context), Dallwitz (1974) and Selkow (1974) suggest groupings for easy identifiability while Vesely (1981) and Michalski (1980 a, b; see also Michalski and Stepp 1985, Stepp and Michalski 1986) both search for clusters which are conceptually easy to characterise. Korhonen (1984) provides a sequential method handling fuzzy assignment and shape variation, while Bezdek and Anderson (19857 and Bezdek *et al.* (1981 a, b) search for lines and planes. Esty (1985) suggests fixing cluster sizes using a negative binomial, Baroni-Urbani and Buser (1976) and Baroni-Urbani (1980) probabilities of occurrence of attributes Switzer (1968) uses generalised loss functions (see also Lefkovitch and Pirktl (1983)), Yamamoto *et al.* (1977) require the consecutive retrieval property, as does Ghosh (1975). Kashyup and Oomen (1983) want measures suitable for sets of strings, Williams, Lance, Webb, Tracey and Dale (1969) for transition matrices using as a target the average matrix. Fukonaga and Flick (1986) test for Gaussian-ness, which also interests Velasco (1980) and Lee (1979).

Sometimes it is not a case of fitting the model to the clusters, but to other things instead. Goodall (1973), Jardine and Sibson (1968) and Hubert and Baker (1977) seek a hierarchy to fit the dissimilarities, but are more restrictive than Sattath and Tversky (1977) who use additive similarity trees in place of ultrametric trees. There are yet more complications when weighting of attributes is also to be calculated, as with Lumelsky's (1982) combined weighting, or Hogeweg's (1976) iterative reweighting also used by Hogeweg and Hesper (1974) to incorporate realignment of sequences, and De Soeto, De Sarbo and Carroll's (1985) alternating least squares weighting.

It is hardly surprising, in these conditions, that definitions of ideal cluster solutions tend to be somewhat vague, as in Demimirmen (1969). To develop an algorithm the ideal cluster must be specified and a fitting procedure found and the great variety of alternatives, all valuable in specific contexts, makes the selection of an appropriate model difficult. Still if a model can be identified and clusters developed to fit data to this model, then adequacy of fit is certainly one way of determining the number of clusters required.

Introvert / extrovert

If you are searching in a hierarchical manner for ideal clusters there is still a further possible means of identifying the right number of clusters, apparently first identified by Williams, Lambert and Lance (1966); the question of level or change. Do you stop when all the potential clusters adequately fit the criterion. Or do you stop when the improvement gained by subdivision is too small to justify the division. This relates to *between* and *within* group variances with appropriate models but can be found in others, *e.g.* simplicity of description in Vesely's predicate calculus definable clusters. Smith and Dubes (1980) comment "a valid cluster starts early in the hierarchy and finishes late". Although they also point out that you could ask for tests of existence of clusters, of the existence of a hierarchy of clusters, or a simple partition, and further for tests of validity - which clusters are real, the point here is that these methods rely on measures of how *compact* and how *isolated* the clusters are in order to judge their validity. In the last section we were primarily concerned with compactness or fit to a model. Many workers, for example Engelman and Hartigan (1969), Hartigan (1978, 1981), Jancey (1974), Mojena (1977), Hartigan (1978), Rogers (1978), Wainer and Schacht (1978), Warnekar and Krishna (1979) and Wong (1984) have emphasised isolation; gaps between clusters, distance between clusters, separability of clusters. Gower's (1974) maximal predictive approach suggests internal predictivity as a criterion but relates this to the degree of erroneous prediction in other groups. However it is often difficult to distinguish this group of approaches from those involving explicit model fitting as is apparent in the studies of Sneath (1980b) and Day (1969b) where distances between clusters are used to test for multivariate normality within.

In some ways it would appear better to attempt to combine the two and compromise a little. Examples of this include Hartigan (1978) who shows that if R is a (function of) Between/Within variances then an approximation to a normal distribution is available, *viz*:

$$\log(1 + R) \sim N\,[1 - \log(1 - 2/P) + 2.4/n\ ,\ 1/(n - 2)]$$

for n entities. Other proposals relating between and within components, not necessarily variances, include Bailey and Cowles (1984), Bock (1985), Feoli and Lausi (1980), Popma *et al.* (1983) and Smith and Dubes (1980). Taken to extremes this leads to a view of science as providing models which fit data adequately for given complexity, and rejects the Popperians "true/false" dichotomy as overly restrictive.

We may add other criteria, such as equitability of numbers and ease of characterisation but these remain *secondary* to the prime theme of balancing the fit we have to the potential gain of generating still more clusters. Cattell (1966) with his "stats" and "aits" actually distinguishes two kinds of cluster, one with good fit, the other with good separation, and it seems likely that visual processing falls here; human observers are very markedly affected by a few extreme points marking boundaries of dense areas. Yoljina and Zagorniko (1978) using human derived criteria seem to end up with a somewhat complex measure of within/between characteristics, coupled with other features such as equitable group sizes.

Dubes and Jain (1976, 1979) have provided an extensive consideration of this approach to establishing the validity of clusters in terms of compactness and isolation. They have presented a number of criteria all of which can aid in the identification of an acceptable number of groups. Cross (1980) uses similar approaches to measuring clustering tendency, while Hogeweg's (1976) iterative reweighting method seems to combine several approaches. It involves a nonparametric test for discriminability, or self consistency but incorporates elements of the complexity/adequacy comparison and of between/within comparison in its use of multiple classifications. Plastria (1986) seeks nonhierarchical clusters but identifies those entities which are "consistently together" and those which are "consistently apart" at all levels in his dual hierarchies, and intersects these two sets to identify natural clusters at several levels! Cliff, McCormick, Zatkin, Cudeck and Collins (1986) employ measures of change in fit in overlapping clusters, although similarity of membership is also a criterion. Orford (1976) also examines various measures as a means of partitioning dendrograms.

All of these methods provide a means of establishing at least the self-consistency of the clusters. They attempt to incorporate intuitively obvious characteristics of clusters in a variety of ways. While this provides extreme flexibility the user of the methods is left in some confusion as to which of these methods provides a useful measuring stick and when. Such evaluative work still remains largely undone.

On talking to angels

Perhaps the most consistent theory of clustering combining coding theory, inductive inference and the notion of compactness and isolation, is that of Wallace and Boulton's (1968) SNOB program. The inductive theory requires balancing adequacy and complexity and is the dominant theme here, though SNOB imposes conditions on the final cluster; for example they must be uncorrelated random collections of points, at least for numeric variables.

Starting from some initial partition, which may be a single group of all the data, the algorithm attempts in the usual suboptimal manner to derive an optimal partition. However this optimal partition is defined in terms of the cost of coding the data so that it may be communicated to some other person. The program can not only shift entities between classes, it can also generate new classes or fuse classes if this reduces the coding cost. In principle, then, it should be able to identify the optimal number of classes. Additionally it can partially assign entities to classes as well since it can identify classes which would suit any entity almost as well as the class in which it is placed. Unfortunately the initial partition seems to be rather constraining, and the suboptimal solutions obtained depend heavily on the start. While choosing more acceptable models for the internal group structure might help somewhat, it seems more likely that a much more extensive, and expensive, search is the only real solution to this problem.

The balancing of adequacy and complexity is of some particular significance in exploratory studies. Breiman *et al.* (1985) use adequacy/complexity for pruning in their predictive classification, although they test the significance of the entire clustering in other ways, such as jackknife testing, and various forms of replication. Cook (1974, see also Cook and Rosenfeld 1976) uses adequacy/complexity arguments in inference procedures coupled with tree structuring to identify acceptable context-free grammars; these are closely related to clustering as can be seen in the studies of Hogeweg and Hesper (1974). Segen and Sanderson (1977) use a similar functional representation but in a search for an adequate Turing machine program! The general notion of compromising between representing the data well and having to operate with a very complex organisation seems attractive. But not to all analysts, of course, for evolutionary taxonomists would not really be content. Such a result would conflict with present notions of the mechanisms of evolution and whatever its merits in information retrieval would remain unacceptable as a representation of the path of evolution.

A la Mode

A geometric view is clearly dominant in those procedures which seek modes, that is clusters of points of high density. Indeed the concept of density of points seems to require an embedding measurable space, although it need not involve a symmetric measure of distance within that space. Modes can be identified in many ways. There are both nonparametric (Hartigan 1985) and parametric (Scott and Thompson 1983, O'Gorman and Sanderson 1984) estimation methods for determining density functions, some of quite exhausting complexity, and in odd spaces (Schaeben 1984). Many of the graph-based methods we shall meet a little later can be regarded as estimating high density clusters (c.f. Hartigan 1981), while the *dip* statistic (Hartigan 1985) is regarded by the author as suitable for testing unimodality.

Other methods include those of Gitman (1973) which first identifies points in areas of locally high density nearest neighbour distances, and then seeks points *central* to these dense points. This seems acceptable except that the number of points needed to surround another increases very rapidly with dimensionality and the method becomes useless in more than 4 or 5 dimensions. In fact many c-means type algorithms also fit here. Examples include Diday and Govaert (1974) and the French "dynamic clustering" school, von Eye (1977), von Eye and Wirsing (1978, 1980), Gavrishin, Coradini and Fulchignoni (1976), Massart, Plastria and Kaufman (1983), Schaeben (1984) using orientation data, Schueler and Wolff (1980), Tou (1979) and Wacker (1972). There are also sequential methods which identify one mode at a time, *e.g.* Uttley (1970) with his sequential "Informon" identifies the most frequent class first.

Some other methods make use of nearest neighbour properties. One example, again related to estimating local probability density functions like Gitman's, is due to Wishart (1969). The method gives a k-neighbour alternative to the usual single linkage methods and has an unusual stopping rule. In the course of processing the number of identified clusters first increases and then decreases the number again. Wishart argues that the maximal number of clusters found is a reasonable guess for the correct number of clusters. The method can of course report a single cluster.

Looking for modes is rather like looking for ideal clusters but avoiding shape constraints to some degree. It also leads to the idea that some entities are peripheral to clusters. This may be good for *discriminating* between clusters, but not for *defining* them. But the notion that not all entities belong equally to a single cluster has itself led to further developments, notably the concept of a fuzzy cluster.

The curate's egg

To Be or Not to Be; that is the question.
To partially be, that is the answer.
Maybe, yet what with statistical draconics and quantal figmatics
we might just get Eric the half-a-bee who exists only temporarily!

With a few exceptions, such as Cliff *et al.* (1986) and some mixture methods, the procedures of clustering emphasise crisp clusters; that is they seek clusters in which every entity is in one, and only one, cluster. In many cases we would like to relax this restriction somewhat. We do not always believe that each entity we are classifying belongs to one class. There is very often a good case to be made for regarding only exceptional entities as having such a property, while the majority of entities are regarded as containing elements of several classes.

There are at least two alternative approaches to obtaining such clusters. In *non-deterministic* clustering an entity can simultaneously belong *equally* to several clusters while in *fuzzy* clustering each entity is ascribed a *degree* of belonging to every cluster. To the first of these classes belong Hartigan's (1972) direct clustering Dale and Anderson's (1973) inosculate clustering Arabie and Carroll's (1980) MAPCLUS, and Cliff *et al.* (1986) overlapping clusters, among others. In these methods the relationship of entity and cluster is usually regarded as reflecting combination of elements of similarity derived from shared subsets of attributes. We have returned once again to the primacy of the description. Although overlap can be derived, as in geographical mapping, from the overlapping of spatial sampling areas and the "real" types, adding the notion of additive classes of attributes, as in inosculate analyses or in additive clustering, would seem to be more fundamental. Perhaps we should look towards the functional similarity of Lewis, Baxendale and Bennett (1967) to provide some answers here. The methods have been little used, largely for computational reasons, but seem to give interesting results where they have been applied.

Jardine and Sibson's (1968) B_k clustering, which allows a fixed number of entities to overlap, seems more closely related to various graph theoretic methods such as clique finding. It emphasises the representation of dissimilarities by the clusters, but higher values of the k parameter seem to lead to excessively complicated results which pose considerable interpretational problems. However because of its emphasis on dissimilarity representation it does not relate closely to the notion of subsets of attributes contributing in combination to the measure. For all that it is perhaps surprising that it has seen so little use since relatively efficient algorithms are available.

Most fuzzy classification methods are related to model-fitting; *e.g.* Bezdek, Windham and Ehrlick (1980), Bezdek, Coray, Gunderson and Watson (1981a, b), Bezdek and Anderson (1985), Gunderson (1982, 1983), Gunderson and Kessel (1978), Windham (1985) all describe c-means algorithms which fit variations on the minimal variance models. Much effort has been made, especially with fuzzy clustering {see *e.g.* Libert and Roubens (1983), Dunn (1974), Ruspini (1969), Roubens (1978, 1982)} to provide means of estimating the number of clusters required. It is interesting that many of the methods proposed use a measure which depends on crispness; the crisper the result the better. This seems to run rather counter to the arguments for using fuzzy groups at all! Most of the criteria do not seem to work too well in practice (*c.f.* McBratney and Moore (1985)), being monotonically related to the number of groups anyway. Some authors have used constrained versions of fuzzy classification, for example Selem and Ismael's (1984) soft clustering. Others have adopted fuzzy paradigms for related problems, for example Ozawa's (1985) use of local density to produce asymmetric similarities, Giles (1976) discussion of fuzzy logics and Nakamura and Iwai's (1982) use of fuzziness in quantifying analogy. Not all uses of fuzzy measures result in fuzzy classes. Gitman (1973) uses fuzzy concepts to identify critical points, at least partially, but does not obtain a fuzzy classification finally.

Instead of seeking fuzzy classes directly, it is possible to first assign entities to crisp classes and then obtain *a posteriori* the fuzzy coefficients of belonging by calculating directly some measure of affinity between entity and cluster; one example would be fuzzy k-neighbour classification. Even more elegant is Korhonen's (1984) learning algorithm which allocates each entity to a class but updates the weights of classes in the same vicinity to emphasise the cluster structure and also to increase separation using what is almost a two-dimensional hash coding system. This points to a relationship with sorting since there are sorting procedures using hash codes applicable to several dimensions.

There is also a question of whether some forms of ordination, *e.g.* multiple group analysis (Thurstone 1945), and perhaps all ordinations, *really* identify ideal types, with fuzzy allocation of entities to these types displayed in a particular manner. That is certainly a possibility worth examining given NoyMeir's (1973) emphasis of unipolar axes and the general attitude of rotating and interpreting axes rather than other features of the display (*c.f.* Guttman's facet analysis for alternatives). But do we have to assume a discontinuous model for classification to tbe regarded as useful?

It's who you know

One very widespread approach to establishing the number and the constituents of clusters relies on graph theoretic representations. By identifying which entities (or nodes) can be regarded as connected by edges, a graph is generated. This graph is then examined to determine if the values of some property of it are significantly different from the values expected if the graph were in fact generated by some random process. There are obviously two questions involved here. First we are concerned with the establishment of the graph representing the relationships between entities which are of interest. Second we must choose which property of random graphs is to be tested to determine if clustering is present.

One obvious method of generating a graph is simply to threshold the similarities, or a subset of them as in Mojena's (1977) proposal, but various other devices have been used for graph-generation purposes. Various forms of single linkage clustering algorithms have been employed, and this may include algorithms which appear to have other objectives, as shown by Shafer Dubes and Jain (1979). The single linkage method, as the minimal spanning tree, underlies both Smith and Jain's (1984) and Panayirci and Dubes' (1983) proposals noted earlier. One advantage of these approaches is the availability of fast algorithms, such as Lehert (1982), but Hartigan (1981) has demonstrated that single linkage has certain desirable consistency properties as a means of estimating modes.

While single linkage has perhaps been most popular, other methods have certainly been tried. Peay (1975) identifies cliques or maximally connected subgraphs which are related to maximal complete subgraphs and complete linkage methods. These also appear in the proposals of Holzner and Stockinger (1973) various k-neighbour and all-neighbour networks such as those employed by Abel and Williams (1981) or Williams and Tracey (1984), or somewhat similar but more mathematically elegant devices for determining associated subsets such as Katajainen and Nevalainen's (1986) relative neighbourhood graphs used by Lefkovitch (1985). Shapiro and Haralick (1979) show that for some special purposes, connectivity between entities can be established directly from observations, without employing explicit similarity measures at all. Futo (1977) has developed an elegant clustering system using multiple connections between entities represented by hypergraph connectivities. He uses binary data and regards the *sharing of a property* as an edge between two entities.

Testing significance has also provided a considerable literature. The properties tested may relate either to the nodes, or to the edges. One of the earliest suggestions using both was provided by Rose (1965), who examined random traversals and identified potential breakpoints from statistics about the frequency of occurrence of an edge or a node on these paths. The edges or nodes most frequently passed through represent ''bridges'' between potential clusters, although Rose provides a test of significance to determine if the frequency is abnormally high.

Most attention has been given to the connectivity within and between clusters, to the number of connected components in a random graph and to the number of connections which reach a particular node. Examples of such approaches include Lingoes and Cooper's (1971) PEPI (probability evaluated partition index), O'Gilvie (1969), Ling (1972, 1973a, b, 1975), Ling and Killough (1976) and Naur and Rabinowitz (1975) on connectivity of random graphs, Burtin (1974) giving asymptotic test values, Schultz and Hubert (1975) and Chen and Fu (1975) on the connectivity of clusters, Matula (1983) using concurrent chaining, and Frank (1978 a, b), Frank and Harary (1982) and Frank and Svenson (1981) exploiting several properties of random graphs. Mojena (1977) has suggested pruning a dendrogram in a hierarchical classification by using the mean and variance of the similarity measures used to construct it, and presumably this could also be applied to more general graphs, although significance testing might prove difficult.

These methods are not without pitfalls, and both Day (1977) and Dubes and Hoffman (1986) suggest some caution in their use. For example, Di Gésu and Maccarone (1986) provide an incorrect solution. Nor are all the methods probabilistic. Wishart's (1969) mode analysis is an extension of single linkage, but the estimate of the number of clusters is not based on a significance test at all, and this is also true of Jancey's (1974) method. Wong and Ghahraman (1980) identify a structural-contextual dichotomy which suggests that several properties of a graph may be interesting simultaneously, and reliance on one feature may be inappropriate.

Methods based on graph-theoretic properties have much attraction, since there is only a very weak model involved. Unfortunately unless the graph is directly observed, then it has to be generated from similarity measures, and these may involve much stronger models. We have to come back to determining what it is meant by *being alike*. Single linkage seems to have both desirable and undesirable properties. It is probably true to say that analysis of any data should probably include the generation of a minimum spanning tree, if at all feasible, since it does provide a variety of interesting information. It is also true that it should not be considered in isolation.

Meaningful relationships

Since many clustering procedures involve the calculation of pairwise similarity measures, there have been many attempts to provide a statistical basis for testing if these measures are significant. If such tests could be found, it is argued, then graphs could be constructed showing entities which are significantly related, and dendrograms could be pruned so that significant similarities existed within groups and/or significant dissimilarities between groups. This is not the only reason for invoking probability arguments in similarity measurement (c.f. Feoli and Lagonegro 1983) and the proposals are not restricted to pairwise similarities. Macnaughton-Smith (1965) has suggested for his information theoretic models, using the change in fit induced by a division into subgroups as a means of terminating an hierarchic division algorithm; he could equally have used the actual fit of the groups. Similar proposals have been made by Ratkowsky and Lance (1978). While useful as heuristics, because the clustering procedure is *maximising* differences ordinary statistical tests are invalid. These proposals are probably best regarded as an improvement on the early practice of chopping dendrograms at some fixed level of similarity. Interestingly the rule can be applied to all groups existing at some point in a classification procedure, or restricted to the context of subdivision of a specific group as suggested by Hill (1980). Thus such stopping rules can be applied in 3 ways; to the individual group, to a pair of potential sibling subgroups, or to all groups at some specific level in the hierarchy. Some other stopping rules have the same potential, notably Sandland and Young's replication test, and the spatial tests of Krishna-Iyer (1949).

Attempts at deriving distributions for similarity coefficients have been reported by many authors, including Rahman (1962), Goodall (1964), Frey (1966), Goodall (1967), Mountford (1971), Toussaint (1974), Tsukamura (1976), Sämdal (1976), Smith and Grassle (1977), Hayes (1978), Stoddard (1979), Strauss (1982), Best, Cameron and Eagleson (1983), Faith (1985), and Lim and Khoo (1985). Most make assumptions about the distributional properties of the data, although Goodall (1964) avoids this necessity since his proposal works in the context of the specific data set. In exploratory analysis it seems unlikely that the assumptions will hold. Sampling is rarely randomised, and often it is difficult to see how it could be. Where the assumptions can be accepted, however, then the tests can be useful.

Most workers do not continue their efforts to the point of describing what to do when the similarities are indeed significant. Obviously a graph could be constructed representing significant relationships between entities, and the procedures described in the previous section employed to partition it. But many clustering methods tend to extremise; they place most similar entities together and separate most different entities. This invalidates assumptions of independence which underlie almost all statistical procedures. As an example Clifford and Goodall (1967) suggested how similarity probabilities might be used in cluster analysis. It is apparent from their discussion that as soon as significance *has* been detected, the basis for the calculation of the similarity measure often disappears and the values can logically be used to cluster the entities only by disregarding the probabilistic basis of their derivation. This paradox appears unresolvable unless quite unrealistic assumptions are made.

The procedure used by Clifford and Goodall also results in many unclustered residuals. Indeed you may decide not to cluster at all! They do not comment on what should then be done to identify patterns in the data. Presumably it would be necessary to use ordination, seriation or techniques derived from spectral analysis although the null hypothesis underlying the similarity measure implies random data and hence no structure of interst. But unclustered residuals can appear at any stage in their analysis, which provides a curious ambivalence.

How to measure significance of similarity with structured and dependent attributes is still largely unknown. Bhapkar and Patterson (1977) have made some suggestions for profiles, Sankoff and Kruskall (1983) for strings, Hubert (1983) for similarity matrices. Lewis, Baxendale and Bennett (1967) for synonymity and Nakamura and Iwai (1982) for analogy. Perhaps it will be possible to employ techniques such as Hogeweg and Hesper's (1984), iterating between tree and similarity measure, in a dynamic re-adjustment of the measure of similarity as the alignment is iteratively *corrected* by use of the classification derived from the previous similarity estimate. Certainly a similar readjustment would be possible in the case examined by Dale, Clifford and Ross (1984) where synonymity was identified by a common context. In this case the context would be adjusted using information from the classification, and then the classification obtained using similarities from the new context.

Feoli and Lagonegro (1983) and Day and Faith (1986) have both examined dualistic similarity. In this case the overall similarity has several components all contributing independently. For example with binary data, similarity is represented by sharing a common state (0/0 or 1/1), dissimilarity by sharing opposed states (0/1 or 1/0). These four components can be combined additively using weights to emphasise specific components. A similar approach is due to Vesely (1981), which leads us to the next section. Significance testing with such dualistic measures has not been studied.

Geometry or Genealogy?

Here we are led to question the whole edifice of polythetic classification and the statistical, geometric and graph theoretic methods which are so widely employed. Instead we seek clusters which represent the results of generating processes, and these clusters need a clear, simple definition. Overall similarity is no longer the most important characteristic of a cluster of entities. Indeed they can be quite disparate overall as long as there is some single important characteristic holding the group together. The important thing in conceptual clustering is the *definitional semantics* of the clusters are regarded as important.

But semantics requires a context for the meaning. A system *generating* the cluster structure should be apparent, and the rules may be intrinsic to the system, as in evolutionary taxonomy, or extrinsic when they are primarily defined by the classifying agent. To define a cluster by some simple rule is also to accept that in the context of the analysis such a simple rule is meaningful. Michalski and Stepp (1985) did not regard ''manufacturer'' as an important attribute of microcomputers, but that is a *value* judgement by them. Their *a priori* assumption the ''primary chip'' (8080, 68000 etc.) was *interesting*, as opposed to other possibilities, led them to accept a classification based on chip type and reject one based on manufacturer. The same criticism applies to their toy trains example (Stepp and Michalski 1986). To identify one particular truck type as interesting because it represents ''carriage of dangerous chemicals'', is to determine a context in which semantic importance can be assessed. Such special applications abound but they *are* special.

This acceptance of special context-dependent objectives has certainly been found profitable in picture processing. Finding lines or curves in pictures, indeed finding any objects in an image, is in many ways just grouping pixels to meet a particular expectation; it is *not* certain that looking for such things will *always* result in interesting outcomes. If I display the abundance values for a species as a grey-scale image, the application of techniques to identify aeroplane shapes is clearly irrelevant; I must look instead for shapes relevant to species distributions, such as the Glass patterns of Phillips and Rosenfeld (1986) perhaps or low-level segmentation approaches such as Trivedi and Bezdek (1986). Analogously, it is often assumed that heuristics from one area of study carry over to another. This is not true and would hardly be expected since heuristics are by definition special sets of rules applicable in particular situations. However when searching for heuristics such practices may themselves represent useful heuristics!

It is also essential to recognise that a simple description may not be possible with the particular set of descriptors employed in the analysis. As an example, let us try to define the concept ''chair''. Chairs are items of furniture on which we sit. This is a functional property. However we can sit on other things such as benches or stools, so that the definition is still imprecise, even ignoring the problem of defining items of furniture. Chairs seem to be constructed to have back-supports, which is a physical property, and this does seem to distinguish them fairly well. But the description is now partly functional, partly physical apparently in some sort of hierarchy of importance. Whether both these kinds of properties would actually be available in any analysis is doubtful. In cladistic studies a parsimony criterion is used as a means of resolution, a means of enforcing a simple description when the required property is not directly recorded, but this leads to very hard computational problems (see Day, Johnson and Sankoff 1986).

Yet the notion of looking for patterns which reflect the generating processes and which are likely to be identifiable in simply defined classes is critical. It probably accounts for the apparent success of quite simple monothetic clustering methods such as Goodall's (1953) clumping method, Williams and Lambert's (1959) "Association Analysis" or Crawford and Wishart's (1967) "Rapid" method. These were later rejected because they did not easily conform to the developing notions of modes and mixtures, and polytheticism. Polythetic methods can be successful in conceptual clustering. Hogeweg and Hesper (1974) provide an actual example of such recovery of generating rules and it is also apparent in the work of Moller-Anderson (1978) and other cladistic taxonomic studies.

This emphasis on simplicity of definition has been articulated most clearly by Michalski (1980 a, b), Michalski and Stepp (1985, but see Dale 1985) and Stepp and Michalski (1986), where it is called "conceptual clustering". It is obvious from some of these studies that the clusters are not defined by properties of all the attributes, but of some select few only. There is a considerable weighting and selection applied, just as in some cladistic studies, for example Farris, Kluge and Eckardt (1970; though not in all c.f. Le Quesne 1974). In these latter only synapomorphies are regarded as relevant to cluster definition. These are defined as shared *derived* states of an attribute so that the user must be able to specify the partial ordering of the states, though the methods concentrate on finding trees rather than clusters.

There are other approaches to generating conceptual clusters. Vesely (1981) adopts a model fitting approach and seeks to define clusters in terms of interesting logical predicates. He maximises the number of such predicates which are true for any cluster, with an ordering of interest in various kinds of predicates. Conjunction, for example, is regarded as more useful than disjunction which is somewhat counter to Michalski's views. Segen and Sanderson (1979) seek a functional definition in a Turing machine generating the patterns, and this may have some relationship both to analogy, which can be treated as a mapping from one structure to another, and to synonymity (c.f. Lewis, Blaxendale and Bennett 1967). It is interesting perhaps that Hogeweg and Hesper (1974) relate their classification to a *grammatical* generating mechanism, specifically a D2L parallel Lindenmeyer system, but inference of grammars, though aided by classification (see *e.g.* Cook and Rosenfeld 1976) is too large a subject to be discussed here. The application of grammars to generating observed species patterns has been studied by Haefner (1978) in some very interesting work, though he is not concerned to derive the grammar automatically.

Conceptual clusters are unlikely to be related to uncorrelated minimal variance properties; we look rather for tightly correlated subsets! This was implicit even in some of the earliest methods of monothetic classification mentioned earlier, for Williams and Lambert (1959) note that the method will find groups where all correlation is indeterminate. It is still to be found in the oligothetic indicators of Hill, Bunce and Shaw's (1975) Indicator species analysis, although the pattern is now very similar to a consecutive retrieval property (*c.f.* Yamamoto *et al.* 1977). This is in spite of the fact that the method is presented as being polythetic! The basic notion is to find necessary and sufficient attributes to simply define the groups of entities. Note that this does NOT conform to the polythetic notion of entities being similar, for in that case any change in *any* attribute is, more or less, equivalent to a change in any other. Indeed it is possible (Dale and Barson unpublished) to obtain polythetic groups such that *no* attribute is shared by more than *two* members. While polythetically acceptable, such groups are extremely difficult to interpret sensibly, unless the pattern of overlapping species is particularly simple. It is so in the seriation model underlying Hill *et al.*'s (1975) algorithm. We are, therefore, not looking for a high density in the space, but for a subspace in which there is a high density or, better, a subspace of low *topological* dimensionality, since there need be no simple means of distinguishing the subspace region.

This is in some ways quite close to the additive clustering model of Arabie and Carroll (1980, see also Shepard and Arabie 1979) which seeks to explain dissimilarity by partitioning it between various sets of attributes. It is alternatively developed in procedures for combining entity and attribute classifications, whether in two steps, as in nodal and inverse analysis (Lambert and Williams 1962) or the later simultaneous procedures of Hartigan (1972) and Dale and Anderson's (1973) inosculate two-parameter analysis. Lambert and Williams (1962) nodal analysis superposes two classifications and then seeks to identify particular entities and attributes which characterise the intersection.

Another method which avoids the minimal variance trap is Gower's maximal predictive analysis (1974, see also Colless 1984). This seeks to identify groups such that within a group the attribute values are *most predictable*. This does of course mean invariance, but in fact results in a tendency to produce indeterminacy, since the attributes ideally have NO variance. It is the confusion of indeterminacy with minimal variance when in fact it provides maximal correlation which has led many workers into the seductions of geometric representations and the pleasures of polytheticism. Note too that we have so far been seeking crisp definitions. Concepts may be fuzzy in the sense that entities, and presumably attributes, are more or less related to them but it does not seem that effective methods exist for finding them. This may not be of great importance. For the purposes of identifying processes generating the data, we need first a clean picture; fuzziness can perhaps intrude later.

Gallimaufry

This section, as its name implies is a heterogeneous collection of methods which seem to have special properties or which do not fit well elsewhere. One possibility lies in using tests for outliers rejecting entities until an acceptable level of homogeneity has been reached. Examples of this approach are seen in Clifford and Goodall's (1967) probabilistic method, and in Wong's feature weighting method which will weight attributes or entities. Other methods can be easily developed. For example we could employ Kendall's (1948) coefficient of concordance to establish a measure of homogeneity of a cluster, coupled with Whitfield's (1952) rank tests to reject oddities. For other tests for outliers see Rohlf (1975) and Hawkins (1979).

Dale and Anderson (1973) and Dale and Webb (1973) have explored the use of models which are symmetric in their use of entities and attributes. This permits the analysis to choose whether one or other should be clustered, and this decision can be used to terminate the clustering process. The two-parameter model employed does permit internal cluster variability to exist; it is not a minimal variance method.

In an isolated study, Murtagh (1983) has attempted to establish probabilistic tests based on hierarchies rather than the clusters implicit in them, that is to determine how likely a particular hierarchy is to occur. It is not obvious when such tests would be valuable, since the topology of the hierarchy does not seem especially interesting. Indeed many results in cladistic taxonomy would suggest that quite unlikely hierarchical forms are not uncommon at all in practice.

Of more significance is the work of Sneath (1985, 1986) which also considers dendrograms, but in this case from the point of view of the distribution of the levels at which fusion occurs. Sneath suggests that, suitably transformed, a hyperspherical multivariate normal distribution of points leads to a normal distribution of levels. Sneath therefore simply tests this normality using a variety of nonparameteric and parametric tests, though the actual test is complicated by variant forms for specific dissimilarity measures and clustering algorithms. Sneath further claims that it is possible to estimate the intrisic dimensionality, that is the effective number of attributes which defines the dimensionality in which the points are embedded, from the cumulative distribution of transformed levels. Although the technique seems to be limited to binary data, this claim is of considerable significance if it proves true, but as yet there is little experience with the proposed test or the estimate. In order to estimate the dimensionality, the user must also supply an estimate of the *intracluster* mean similarity. Sneath (1980b) has also looked at the problem of recognition of clusters in low-dimensional ordination diagrams, a practice which is widespread and in my own experience is often effective enough.

Another somewhat isolated approach is that of Warnekar and Krishna (1979). Their objective is to identify clusters which are linearly separable. Such a property would of course be extremely useful, but does seem rather restrictive as a general aim. However ease of discrimination is certainly a property of conceptual clusters, though Warnekar and Krishna are not really concerned with logical definition so much as numerical.

Finally, Binder (1978, 1981) has explored the possibilities of using Bayesian methods in clustering. This involves the estimation of *a priori* probabilities and presently seems most closely related to the minimal variance clustering approaches. He does comment that the number of clusters could simply be a extra parameter in the estimation procedure but the computational difficulties seems to be rather large. More experience with such methods is certainly needed.

Though isolated from the mainstream of developments for significance testing in cluster analysis, it should be clear that the methods discussed in this section have considerable potential. It seems unlikely that any great breakthrough is likely in Bayesian methods whatever their theoretical advantages, but the empirical results of Sneath, in particular, suggest that simple testing for clusters may not be unobtainable.

The choice is yours

In the Introduction it was argued that an examination of kinds of stopping rules would lead us to some better understanding of the nature of clustering and the choices available to the user of clustering methods. Hopefully at the end of the examination we would have a clearer idea of the questions to be asked when choosing a clustering method. Then, having decided how we can organise our stopping rules, we can look to see in what situations one or other of them might be preferable. And perhaps we shall really know when to stop or at least what to do to make a sensible decision. Have these aims been attained?

What has become clear is that, as Gasking (1960) indicated, no one model for clustering will be adequate in all circumstances. The agent using clustering will find it necessary to decide what he wishes the clustering to do. This does not mean that clusters produced by some other method will not be interpretable, useful or even significant. There is too much overlap in the objectives and algorithms. But the clusters will be conceptually significantly less useful than they could be if the apposite method were used.

It might be possible to proceed by first determining if you are *really* looking for clusters based on a particular pattern of attribute values, not looking for some organisation which will aid your understanding of the generating process. The latter will also commonly imply an interest in hierarchy. Sometimes the sampling will dominate the choice of method. If you have only sampled relationships between entities, then you are not easily able to model the underlying population features. Relationships can of course be asymmetric, which itself might limit the choice of approach to graph theoretic models. It may be that an ordination will suit better the objectives of the study; ordination has problems of its own of course, and seriation rather worse ones but that is another story.

So we seem to have a variety of objectives; looking for modes, looking for collections of points which fit some particular pattern (lines, circles), can be sorted in particular ways, whose interpoint distances can be represented different ways (trees, additive clusters, ordination) which belong to the unspecified types in differing degrees (fuzzy) or uniquely (crisp), which can be coded in simple ways or finally which aid the user in identification of groups. Sometimes classification is simply one step in a further study. When it is used as one means of developing an alphabet for examination of sequences, for example, it is reducing a multivariate data set to a single multistate variable (Dale 1979). Part of the problem is that the nature of a cluster definition is related also to the sampling of some universe of discourse, to description and to the nature of similarity.

One major class of clustering approaches regards clusters as modes in a general sense of high density regions perhaps of some specified shape. This shape has most commonly been multivariate normal without correlation which is a most unlikely situation. With clusters as collections corresponding more or less adequately, and with greater or lesser complexity, to some prescribed pattern of variation the modal idea is extended and this is in part true for graph-based clusters of connectivities. Often it is the parameters of the underlying model which are important here as descriptors of the clusters, and the descriptions can be complex.

In contrast to these is the conceptual approach, where clusters represent specific outcomes of some generation process, where correlation is sought not rejected and the objective is really the inferences which may be drawn concerning the processes underlying the data. Here it is how the cluster came to be which is important. It is the functioning of the generating processes which is the major interest. As noted above these two results obtained by using these approaches need not be disjoint but they surely represent different functions.

Whatever the final solution adopted in choosing the clustering method, the paper does show that a variety of methods are available for establishing the validity of clusters. Many of these are rather specific in their requirements, and so may not be universally applicable. However the major problem would seem to be the lack of available computer programs to explore the efficacy of the proposals. The TAXON package (Ross 1979) now contains some validity tests as part of a general classification and ordination package. Users of the package, when they become aware of these particular tests do find them useful and actively examine their data to see if any of the tests are applicable (*c.f.* Dale *et al.* 1984). There is certainly much work necessary to improve the present tests and to supply novel ones, there is, unfortunately, small sign that clustering packages are moving to include more tests of validity.

Acknowledgement. I would like to thank Prof. Enrico Feoli for his constant enthusiasm and encourangement, while I was in Trieste, and the other members of the Departimento di Biologia who suffered me without complaint, Mario Lagonegro and especially Marina Geatti.

REFERENCES

ABEL, D.J. and W.T. WILLIAMS. 1981. NEBALL and FINGRP: new programs for multiple nearest neighbour analysis. Austral. Comput. J. 13: 24-26.

ARABIE, P. and J.D. CARROLL. 1980. MAPCLUS: a mathematical programming approach to fitting the ADCLUS model. Psychometrika 45: 211-235.

BACKER, E. 1978. *Cluster analysis by optimal decomposition of induced fuzzy sets.* Delft Univ. Press, pps 235.

BAILEY, T. and J. COWLES. 1984. Cluster definition by optimization of a simple measure. IEEE Trans. Patt. Anal. Mach. Intel. PAMI-6: 645-652.

BARONI-URBANI, C. 1980. A statistical table for the degree of coexistence between two species. Oecologia (Berl) 44: 287-289.

BARONI-URBANI, C. and H.W. BUSER. 1976. Similarity of binary data. Syst. Zool. 25: 251-259.

BASFORD, K. and G.J. MCLACHLAN. 1985a. Estimation of allocation rates in a cluster analysis context. J. Amer. Statist. Assoc. 80: 286-293.

BASFORD, K. and G.J. MCLACHLAN. 1985b. The mixture method of clustering applied to three-way data. J. Classif. 2: 109-125.

BEST, D.J., M.A. CAMERON and J.K. EAGLESON. 1983. A test for comparing large sets of tau values. Biometrika 70: 447-453.

BEZDEK, J.C. and I.A. ANDERSON. 1985. An application of the c-varieties clustering algorithms to polygonal curve fitting. IEEE Trans. Systems, Man and Cybernetics SMC-15: 637-641.

BEZDEK, J.C., C. CORAY, R. GUNDERSON and J. WATSON. 1981a. Detection and characterization of cluster substructure I. linear structure: fuzzy c-lines. SIAM J. Appl. Math. 40:

339-371.

Bezdek, J.C., C. Coray, R. Gunderson and J. Watson. 1981b. Detection and characterization of cluster substructure II. Fuzzy c-varieties and convex combinations thereof SIAM J. Appl. Math. 40: 358-372.

Bezdek, J.C. M.P. Windham and R. Ehrlick, 1980. Statistical parameters of cluster validity functionals. Intern. J. Comput. Inform. Sci. 9: 323-336.

Bhapkar, V.P. and K.W. Patterson. 1977. On some nonparametric tests for profile analysis of several multivariate samples. J. Multivar. Anal. 7: 265-273.

Binder, D.A. 1978. Bayesian cluster analysis. Biometrika 65: 31-38.

Binder, D.A. 1981. Approximations to Bayesian clustering rules. Biometrika 68: 275-285.

Bock, H.H. 1985. On some significance tests in cluster analysis. J. Classif. 2: 77-108.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone. 1984. *''Classification and Regression Trees''*. Wordsworth, Belmont, Ca.

Burtin, Yu.D. 1974. On extreme metric parameters of a random graph I. Asymptotic estimates. Theory Probab. Appl. 19: 710-725.

Cattell, R.B. and M.A. Coulter. 1966. Principles of behavioural taxonomy and the mathematical basis of the TAXONOME computer program. Brit. J. Math. Statist. Phychol. 19: 237-269.

Česka, A. and H. Roemer. 1971. A computer program for identifying species-relevé groups in vegetation studies. Vegetatio 23: 255-276.

Chen, Z. and K.-S. Fu. 1975. On the connectivity of clusters. Inform. Sci. 8: 283-299.

Chiu, D.K.Y. and A.K.C. Wong. 1986. Synthesizing knowledge: a cluster analysis approach using event covering. IEEE Trans. Syst. Man Cybern. SMC-16: 251-259.

Cliff, N., D.J. McCormick, J.L. Zatkin, R.A. Cudeck and L.M. Collins. 1986. BINCLUS: nonhierarchical clustering of binary data. Multivar. Behav. Res. 21: 201-227.

Clifford, H.T. and D.W. Goodall. 1967. A numerical contribution to the classification of the Poaceae. Austral. J. Bot. 15: 499-519.

Cohen, V. and J. Obadia. 1974. Inverse data analysis COMPSTAT 1974. pp 141-148.

Cole, A.J. and D. Wishart. 1970. An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. Comput. J. 13: 156-163.

Colless, D.H. 1984. A method for hierarchical clustering based on predictivity. Syst. Zool. 33: 64-68.

Cook, C.M. 1974. *Grammatical Inference by Heuristic Search*. Dept. Comput. Sci., Univ. Maryland, College Park, Maryland. Rep. TR-287. 109 pps.

Cook, C.M. and A. Rosenfeld. 1976. Some experiments in grammatical inference. in: J.C. Simon (ed.) Computer Oriented Learning Processes. Nordhoolt, Leiden. pps 157-174.

Crawford, R.M.M. and D. Wishart. 1967. A rapid multivariate method for the detection and classification of groups of ecologically related species. J. Ecol. 55: 505-524.

Cross, G. 1980. Some approaches to measuring clustering tendency. Dept. Comput. Sci., Coll. Engng, Michigan State Univ. Tech. Rep. TR-80-03. pps. 69.

Dale, M.B. 1979. On linguistic approaches to ecosystems and their classification. In: *Multivariate Methods in Ecological Work* L. Orlóci, C.R. Rao and M.W. Stiteler (eds.) Statistical Ecology ser. 7. pps. 11-20. Internatl. Coop. Publish. House, Maryland.

Dale, M.B. 1985. On the comparison of conceptual clustering and numerical taxonomy. IEEE Trans. Patt. Anal. Mach. Intel. PAMI-7: 241-244.

Dale, M.B. and D.J. Anderson. 1973. Inosculate analysis of vegetation data. Austral. J. Bot. 21: 253-276.

Dale, M.B., H.T. Clifford and D.R. Ross. 1984. Species, equivalence and morphological redescription: a Stradbroke Island vegetation study. In: R.J. Coleman, J. Covacevich and P. Davie (eds.) *Focus on Stradbroke: New Information on North Stradbroke Island and surrounding areas, 1974-1984*. Boolarong Publ., Brisbane and Stradbroke Island Management Organization, Amity Point.

Dale, M.B. and D. Walker, 1970. Information analysis of pollen diagrams. Pollen et Spores 12: 21-37.

Dale, M.B. and L.J. Webb. 1975. Numerical methods for the establishment of Associations. Vegetatio 30: 77-87.

Dale, P.E.R., K. Hulsman, B.R. Jahnke and M.B. Dale. 1984. Vegetation and nesting preferences of black noddies at Masthead Island., Great Barrier Reef. I. Patterns at the macro scale. Austral. J. Ecol. 9: 335-341.

Dallwitz, M.J. 1974. A flexible computer program for generating identification keys. Syst. Zool. 23: 50-57.

D'Andrade, R.G. 1978. U-statistic hierarchical clustering. Psychometrika 43: 59-67.

Davis, B.R. 1985. An associative hierarchical self-organising system. IEEE Trans. Systems Man and Cybernetics SMC-15: 570-579.

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

Day, N.E. 1969b. Divisive cluster analysis and a test for multivariate normality. Internatl. Statist. Inst. Bull. 43: 110-112.

Day, W.H.E. 1977. Validity of clusters formed by graph theoretic methods. Math. Bio Sci. 36: 299-317.

Day, W.H.E. and D.P. Faith. 1986. A model in partial orders for comparing objects by dualistic measures. Math. Bio Sci. 78: 179-192.

Demimirmen, F. 1969. *Multivariate procedures and* FORTRAN IV *programs for evaluation and improvement of classification*. Kansas Geolog. Surv. Comput. Contrbtn. 31. pps 51.

De Soete, G., W.S. de Sarbo and J.D. Carroll. 1985. Optimal variable weighting for hierarchical clustering: an alternating least squares algorithm. J. Classif. 2: 173-192.

Diday, E. and G. Govaert. 1974. Classification avec distance adaptive. C.R. Acad. Sci. Paris, A 993-995.

Di Gesù, V. and M.C. Maccarone. 1986. Feature selection and 'possibility theory' Patt. Recog. 19: 63-72.

Dubes, R.C. and A.K. Jain. 1976. Clustering techniques: the user's dilemma. Patt. Recog. 8: 247-260.

Dubes, R. and A.K. Jain. 1979. Validity studies in clustering methodologies. Patt. Recog. 11: 235-254.

Dubes, R. and A.K. Jain. 1980. Clustering methodologies in exploratory data analysis. Adv. Comput. 19: 113-228.

Dubes, R.C. and R.L. Hoffman. 1986. Remarks on some statistical properties of the minimum spanning forest. Patt. Recog. 19: 49-53.

Dunn, J.C. 1974. A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters. J. Cybernet. 3: 22-57.

Ecob, R. 1978. An empirical evalutation of the behaviour of selected measures of tree and partition similarity in relation to the investigating of the sampling statistics of AID. Egyptian Statist. J. 22: 1-27.

EDELBROCK, C. 1979. Mixture model tests of hierarchical clustering algorithms: the problem of classifying everybody. Multiv. Behav. Res. 14: 367-384.

EIGEN, D.J., R.F. FROMM and R.A. NORTHOUSE. 1974. Cluster analysis based on dimensional information with application to feature selection and classification. IEEE Trans. Systems Man and Cybernetics SMC-4: 284-294.

ENGELMAN, L. and J.A. HARTIGAN. 1969. Percentage points of a test for clusters. Amer. Statist. Assoc. J. 64: 1647-1648.

ESTY, W.W. 1985. Estimation of the number of classes in a population and the coverage of a sample. Math. Scientist. 10: 41-50.

EYE, A. von 1977. Über die Verwendung von Quadriken zur einbeschreibenden Klassifikation. Biom. J. 19: 283-290.

EYE, A. von and M. WIRSING. 1978. An attempt for a mathematical foundation and evaluation of MACS, a method for multidimensional automatical cluster detection. Biom. J. 20: 655-666.

EYE, A. von and M. WIRSING. 1980. Cluster search by enveloping space density maxima. COMPSTAT 1980, Physica-verlag, Vienna, pp 447-45.

FAITH, D.P. 1985. A model of immunological distance in systematics. J. theor. Biol. 114: 511-526.

FARRIS, J.S., A.G. KLUGE and M.J. ECKARDT. 1970. A numerical approach to phylogenetic systematics. Syst. Zool. 19: 172-189.

FELSENSTEIN, J. 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. 14: 313-333.

FEOLI, E. and M. LAGONEGRO. 1983. A resemblance function based on probability: applications to field and simulated data. Vegetatio 53: 3-9.

FEOLI, E. and M. LAGONEGRO. 1984. Effects of sampling intensity and random noise on detection of species groups by intersection analysis. Studia Geobotanica 4: 101-108.

FEOLI, E. and D. LAUSI. 1980. Hierarchical levels in syntaxonomy based on information functions. Vegetatio 42: 113-115.

FRANK, O. 1978a. *Inferences concerning cluster structure*. Dept. Statistics, Univ. Lund, CODEN: LUSADG/STAT-3050/1-7.

FRANK, O. 1978b. Estimation of the number of connected components in a graph by using a sampled subgraph. Scand. J. Statist. 5: 177-188.

FRANK, O. and F. HARARY. 1982. Cluster inference by using transitivity indices in empirical graphs. Amer. Statist. Assoc. J. 77: 835-840.

FRANK, O. and K. SVENSSON. 1981. On probability distributions of single linkage dendrograms. J. Statist. Comput. Simul. 12: 121-131.

FREY, T. 1966. On the significance of Czekanowki's index of similarity. Applicationes Mathematicae 9: 1-7.

FRID, L.M. 1970. Minimization of a function specified over a tree. Kybernetika 4: 115-119.

FRIEDMAN, J. and L.C. RAFSKY. 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Statist. 7: 697-717.

FUKUNAGA, K. and T.E. FLICK. 1986. *A test of the Gaussianness of a data set using clustering*. IEEE Trans. Patt. Anal. Mach. Intel. PAMI-8: 240-247.

FUTO, P. 1977. A new model and algorithm for cluster analysis. Szigma 10: 199-220.

GANASALINGAM, S. and G.J. MCLACHLAN. 1979. A case study of two clustering methods based on maximum likelihood. Statist. Neerland. 33: 81-90.

GANESALINGAM, S. and G.J. MCLACHLAN. 1980. A comparison of mixture and classification approaches to cluster analysis. Commun. Statist.-Theor. Meth. A9: 923-933.

GASKING, D. 1960. Clusters. Australas. J. Phil. 38: 1-36.

GAVRISHIN, A.I., A. CORADINI, and M. FULCHIGNONI. 1976. *On the formulation of the new z^2 criterion*. Lab. Astrofisica Spaziale, Rap. 19, Frascati.

GHOSH, S.P. 1975. Consecutive storage of relevant records with redundancy. Commun. A.C.M. 18: 464-471.

GILBERT, N. and T.C.E. WELLS. 1966. The analysis of quadrat data. J. Ecol. 54: 675-685.

GILES, R. 1976. Lukasiewicz logic and fuzzy set theory. Int. J. Man-Machine Stud. 8: 313-327.

GITMAN, I. 1973. An algorithm for nonsupervised pattern classification. IEEE Trans. Systems Man and Cybernetics SMC-3: 66-74.

GOLDEN, R.R. and P.E. MEEHL. 1980. Detection of biological sex: an empirical test of cluster methods. Multiv. Behav. Res. 15: 475-496.

GOODAL, D.W. 1953. Objective methods for the classification of vegetation I. The use of positive interspecific correlation. Austral. J. Bot. 1: 39-63.

GOODALL, D.W. 1964. A probabilistic similarity index. Nature 203-1098.

GOODALL, D.W. 1967. The distribution of the matching coefficient. Biometrics 23: 647-656.

GOODALL, D.W. 1969. A procedure for the recognition of uncommon species combinations in sets of vegetation samples. Vegetatio 18: 19-35.

GOODALL, D.W. 1973. Sampling similarity and species correlation. In: R.H. Whittaker (ed.), *Handbook of Vegetation Science*, Vol. 5, pps 105-156. Junk, The Hague.

GORDESCH, J. and P.P. SINT. 1974. Clustering structures. COMPSTAT 74. 82-92.

GOTOH, O. 1986. Alignment of three biological sequences with an efficient traceback procedure. J. theoret. Biol. 121: 327-337.

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

GOWER, J.C. and C.F. BANFIELD. 1978. Goodness of fit criteria for hierarchic classification and their empirical functions. Proc. 8th Internatl. Biometrics Symp. Constanz. pps. 347-361.

GUNDERSON, R.W. 1982. Choosing the r-dimension for the FCV family of clustering algorithms. BIT 22: 140-149.

GUNDERSON, R.W. 1983. An adaptive FCV clustering algorithm. Interntl. J. Man-Mach. Stud. 19: 97-104.

GUSTAFSON, D.E. and W.E. KESSEL. 1978. Fuzzy clustering with a fuzzy covariance matrix. In: D.S. Fu (ed.) Proc. IEEE Conf. Decision Control. pps. 761-76.

HAEFNER, J.W. 1978. Ecosystem assembly grammars: generative capacity and empirical adequacy. J. theor. Biol 73: 293-318.

HÁJEK, P. and T. HAVRÁNEK. 1978. The GUHA method - its aims and techniques (twenty-four questions and answers). Int. J. Man Mach. Stud. 10: 3-22.

HARPER, C.W. Jr. 1978. Groupings by locality in community ecology and palaeoecology. Lethaia 11: 251-257.

HARTIGAN, J. 1972. Direct clustering of a data matrix. Amer. Statist. Assoc. J. 67: 123-129.

HARTIGAN, J. 1978. Asymptotic distribution of a clustering criterion. Ann. Statist. 6: 117-131.

HARTIGAN, J. 1981. Consistency of single linkage for high density clusters. Amer. Statist. Assoc. 76: 388-396.

HARTIGAN, J.A. 1985. Statistical theory in clustering. J. Classif. 2: 63-76.

HARTIGAN, P. 1985. Algorithm AS 217. Computation of the Dip statistic to test for unimodality. Appl. Stat. 34: 320-325.

HAWKINS, D.M. 1979. Fractiles of an extended multiple outlier test. J. Statist. Comput. Simul. 8: 227-336.

HAYES, W.B. 1978. Some sampling properties of the Fager index for recurrent species groups. Ecology 59: 194-196.

HILL, M.O., R.G.H. BUNCE and M.W. SHAW. 1975. Indicator species analysis, a divisive polythetic method of classification and its application to a survey of native pine-woods in Scotland. J. Ecol. 63: 597-613.

HILL, R.S. 1980. A stopping rule for partitioning dendrograms. Bot. Gaz. 141: 321-324.

HOGEWEG, P. 1976. Iterative character weighting in numerical taxonomy. Comput. Biol. Med. 6: 199-211.

HOGEWEG, P. and B. HESPER. 1974. A model study of biomorphological description. Patt. Recog. 6: 165-179.

HOGEWEG, P. and B. HESPER. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J. Mol. Evol. 20: 175-186.

HOLZNER, W. and F. STOCKINGER. 1973. *Der Einsatz von Elektonenrechnern bei der pflanzensoziologischen Tabellenarbeit*. Österr. Bot. Z. 121: 303-309.

HSU, Y-S., J.J. WALKER and D.E. OGREN. 1986. A stepwise method for determining the number of component distributions in a mixture. Math. Geol. 18: 153-160.

HUBERT, L.J. 1983. Inference procedures for the evaluation and comparison of proximity matrices. In; J. Felsenstein (ed.), *Numerical Taxonomy*, pps. 209-228. Springer-Verlag. Berlin.

HUBERT, L.J. and P. ARABIE. 1985. Comparing partitions. J. Classif. 2: 193-218.

HUBERT, L.J. and F.B. BAKER. 1977. An empirical comparison of baseline models for goodness-of-fit in r-diameter hierarchical clustering. In: J. van Ryzin (ed.), *Classification and Clustering*, pps. 131-151. Academic Press, New York.

HUXLEY, A. 1937. "Ends and Means." (An enquiry into the nature of ideals and into the methods employed for their realization.) Chatto and Windis, London.

JACKSON, D.M. 1970. The stability of classifications of binary data. Classif. Soc. Bull. 2: 44-46.

JACKSON, D.M. 1972. Stability problems in non-statistical classification theory. Comput. J. 15: 214-221.

JAIN, N.C. A. INDRAYAN and L.R. GOEL. 1986. Monte Carlo comparisons of six hierarchical clustering methods on random data. Patt. Recog. 19: 95-99.

JANCEY, R.C. 1974. Algorithm for the detection of discontinuities in data sets. Vegetatio 29: 131-133.

JARDINE, N. and R. SIBSON. 1968. The construction of hierarchic and nonhierarchic classifications. Comput. J. 11: 177-184.

KASHYAP, R.L. and B.J. OOMMEN. 1983. Similarity measures for sets of strings. Intern. J. Comput. Math. 13: 95-104.

KATAJAINEN, J. and O. NEVALAINEN. 1986. Computing relative neighbourhood graphs in the plane. Patt. Recog. 19: 221-228.

KENDALL, M.G. 1948. *Rank Correlation Methods*. Griffin, London.

KLAUBER, M.R. 1975. Space-time clustering test for more than two samples. Biometrics 31: 719-726.

KLOPMAN, G. and O.T. MACINA. 1985. Use of the computer automated structure evaluation program in determining quantitative structure-activity relationships with hallucinogenic phenylalkylamines. J. theor. Biol. 113: 637-648.

KORHONEN, T. 1984. *Self-Organization and Associative Memory*. pps 125-188. Springer-Verlag, Berlin.

KRISHNA-IYER, P.V. 1949. The first and second moments of some probability distributions arising from points on a lattice and their application. Biometrika 36: 135-141.

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

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

LANCE, G.N. 1970. Mixed and discontinuous data. In: R.S. Anderssen and M.R. Osborne (eds.), *Data Representation*, pps. 102-107. Univ. Queensland Press, St. Lucia, Qld.

LANCE, G.N. and W.T. WILLIAMS. 1977. Attribute contributions to a classification. Austral. Comput. J. 9: 128-129.

LANGRIDGE, D.J. 1971. *On the Computation of Shape*. Intern. Conf. Frontiers Patt. Recog. 35 pps. Honolulu, Hawai.

LE QUESNE, W.J. 1974. The uniquely derived character concept and its cladistic application. Syst. Zool. 23: 513-517.

LEE, K.L. 1979. Multivariate tests for clusters. Amer. Statist. Assoc. J. 74: 708-714.

LEE, R.C.T., J.R. SLAGLE and C.T. MONG. 1976. *Application of Clustering to Estimate missing Values and Improve Data Integrity*. Proc. 2nd Internatl. Conf. Software Engrng, San Francisco. pps 539-544.

LEFKOVITCH, L.P. 1975. Choosing clustering levels for nonhierarchical procedures. In: G.F. Estabrook (ed.), *Proc. 8-th Internatl. Conf. Numerical Taxonomy*, pps. 132-142.

LEFKOVITCH, L.P. 1976. A loss function minimization strategy for grouping from dendrograms. Syst. Zool. 25: 41-48.

LEFKOWITCH, L.P. 1978. Cluster generation and grouping using mathematical programming. Math. Bio Sci. 41: 91-110.

LEFKOVITCH, L.P. 1980. Conditional clustering. Biometrics 36: 43-58.

LEFKOVITCH, L.P. 1982. Conditional clusters, musters and probability. Math. Bio Sci. 60: 207-234.

LEFKOVITH, L.P. 1985. Further nonparametric tests for comparing dissimilarity matrices based on the relative neighbourhood graph. Math. Bio Sci. 73: 71-88.

LEHERT, P. 1982. Clustering by connected components in O (n) expected time. RAIRO Informat. 15: 207-218.

LENNINGTON, R.K. and R.H. FLAKE. 1975. Statistical evaluation of a family of clustering methods. In: G.F. Estabrook (ed.), *Proc. 8-th Interntl. Conf. Numerical Taxonomy*, pps. 1-37.

LESSIG, V.P. 1972. Comparing cluster analyses with cophenetic correlation. J. Marketing Res. 9: 82-84.

LEWIS, P.A.W., P.B. BAXENDALE and J.L. BENNET. 1967. Statistical discrimination of the Synonymy/Antonymy relationship between words. Assoc. Comput. Mach. J. 14: 20-44.

LIM, T.M. and H.W. KHOO. 1985. Sampling properties of Gower's general coefficient of similarity. Ecology 66: 1682-1685.

LING, R.F. 1972. On the theory and construction of k-clusters. Comput. J. 15: 326-332.

LING, R.F. 1973a. The expected number of components in random linear graphs. Ann. Probab. 1: 876-881.

LING, R.F. 1973b. A probability theory of cluster analysis. Amer. Statist. Assoc. J. 68: 159-154.

LING, R.F. 1975. An exact probability distribution on the connectivity of graphs. J. Math. Psychol. 12: 90-96.

LING, R.A. and G.C. KILLOUGH. 1976. Probability tables for cluster analysis based on a theory of random graphs. Amer. Statist. Assoc. J. 71: 293-300.

LINGOES, J. and T. COOPER. 1971. PEP-I. A FORTRAN IV (G) program for Guttman-Lingoes nonmetric probability clustering. Behav. Sci. 16: 259-261.

LIBERT, G. and M. ROUBENS. 1983. New experimental results in cluster validity of fuzzy clustering algorithms. In: J. Janssen, J-F. Marcotorchino and J-M. Proth (eds.), *New Trends in Data Analysis and Applications*. pps. 205-218. Elsevier, North Holland.

LITTLE, I.P. and D.R. ROSS. 1985. The Levenshtein metric, a new means for soil classification tested by data from a sand-podzol chronosequence and evaluated by discriminant analysis. Aust. J. Soil. Res. 23: 115-130.

LÓPEZ DE MÀNTARAS, R. and J. AGUILAR-MARTIN. 1985. Self-learning pattern classification using a sequential clustering technique. Patt. Recog. 18: 271-277.

LUKASOVÁ, A. 1979. Hierarchical agglomerative clustering procedure. Patt. Recog. 11: 365-381.

LUMELSKY, V.J. 1982. A combined algorithm for weighting the variables and clustering in the clustering problem. Patt. Recog. 15: 53-60.

MACNAUGHTON-SMITH, P. 1965. *Some Statistical and Other Numerical Methods for Classifying Individuals*. Home Office Res. Unit. Rep. 6. 65 pps.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression. Canc. Res. 27: 209-220.

MARGULES, C.R., D.P. FAITH and L. BELBIN. 1985. An adjacency constraint in agglomerative hierarchical classifications of geographic data. Environ. Planning A-17: 397-412.

MASSART, D.L., F. PLASTRIA and L. KAUFMAN. 1983. Non-hierarchical clustering with MASLOC. Patt. Recog. 16: 507-516.

MATULA, D.W. 1983. Cluster validity by concurrent chaining. In: J. Felsenstein (ed.) *Numerical Taxonomy*, pps. 156-166. Springer Verlag, Berlin.

MCBRATNEY, A.R. and A.W. MOORE. 1985. Application of fuzzy sets to climatic classifications. Agric. For. Meteor. 35: 165-185.

MICHALSKI, R.S. 1980a. Pattern recognition as rule-guided inductive inference. IEEE. Trans. Patt. Anal. Mach. Intel. PAMI-2: 349-361.

MICHALSKI, R.S. 1980b. Knowledge acquisition through conceptual clustering: a theoretical framework and an algorithm for partitioning data into conjunctive concepts. J. Policy Anal. Inform. Sci. 4: 219-244.

MICHALSKI, S. and R.E. STEPP. 1985. Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Trans. Patt. Anal. Mach Intel. PAMI-5: 396-410.

MICHAUD, P. 1983. Opinions aggregation. In: J. Janssen, J-P Marcotorchino and J-M. Proth (eds.), *New Trends in Data Analysis and Applications*, pp. 5-27 Elsevier, North Holland.

MILLIGAN, G.W. 1981. A review of Monte Carlo tests for clustering. Multiv. Behav. Res. 16: 379-407.

MILLIGAN, G.W. and P.D. ISAAC. 1980. The validation of four ultrametric clustering algorithms. Patt. Recog. 13: 41-50.

MILLIGAN, G.W. and V. MAHAJAN. 1980. A note on procedures for testing the quality of a clustering of a set of objects. Decis. Sci. 11: 669-677.

MILLIGAN, G.W., S.C. SOON and L.M. SOKOL. 1983. The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure. IEEE Trans. Patt. Anal. Mach. Intel. PAMI-5: 40-47.

MINKOFF, E.C. 1965. The effect on classification of slight alterations in numerical taxonomy. Syst. Zool. 15: 196-213.

MOJENA, R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. Comput. J. 20: 359-363.

MOLANDER, P. 1986. Induction of categories: the problem of multiple equilibria. J. Math. Psychol. 30: 42-54.

MOLLER-ANDERSON, N. 1978. Some principles and methods of cladistic analysis with notes on the uses of cladistics in classification and biogeography. Z. Zool. Syst. Evolutionsforsch. 16: 242-255.

MOUNTFORD, M.D. 1971. A test of the difference between two clusters. In: Patil, G.P., Pielou, E.C. and Waters, W.E. "Statistical Ecology 3." Penn. State Univ. Press pp 237-251.

MURTAGH, F. 1983. A probability theory of hierarchic clustering using random dendrograms. J. Statist. Comput. Simul. 18: 145-157.

NAKAMURA, K. and S. IWAI. 1982. A representation of analogical inference by fuzzy sets and its application to information retrieval system. In: M.M. Gupta and E. Sanchez (eds.), *Fuzzy Information and Decision Processes*, pps. 373-386. North Holland, Amsterdam.

NAUR, J.I. and L. RABINOWITZ. 1975. The expectation and variance of the number of components in random linear graphs. Ann. Probab. 3: 159-161.

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

O'CALLAGHAN, J.F. 1976. A model for recovering perceptual organization from dot patterns. IEEE 3-rd Internatl. Conf. Patt. Recog. Proc. pp 294-298.

O'GILVIE, J.C. 1969. The distribution of number and size of connected components in a random graphs of medium size. Information Processing 68, North Holland, Amsterdam. pp 1527-1530.

O'GORMAN, L. and A.C. SANDERSON. 1984. The converging squares algorithm: an efficient method for locating peaks in multidimensions. IEEE. Trans. Patt. Anal. Mach. Intel. PAMI-6: 280-288.

ORFORD, J.D. 1976. Implementation of criteria for partitioning a dendrogram. Math. Geol. 8: 75-84.

OZAWA, K. 1983. CLASSIC: a hierarchical clustering algorithm based on asymmetric similarities. Patt. Recog. 16: 201-211.

OZAWA, K. 1985. A stratificational overlapping cluster scheme. Patt. Recog. 18: 279-286.

PALKA, Z. 1982. Isolated trees on a random graph. Zastos. Matem. 17: 309-316.

PANAYIRCI, E. and R.C. DUBES. 1983. A test for multidimensional clustering tendency. Patt. Recog. 16: 433-444.

PAWLAK, Z. 1984. Rough classification. Int. J. Man-Mach. Studies 20: 469-483.

PEAY, E.H. 1975. Nonmetric grouping: Clusters and cliques.

Psychometrika 40: 297-313.

Perillo, G.M.E. and E. Marone. 1986a. Determining optimal numners of class intervals using maximal entropy. Math. Geol. 18: 401-407.

Perillo, G.M.E. and E. Marone. 1986b. Applications of the maximal entropy and optimal number of class interval concept: two examples. Math. Geol. 18: 465-475.

Phillips, T.H. and A. Rosenfeld. 1986. A simplified method of detecting structure in Glass patterns. Patt. Recog. Lett 4: 213-217.

Pirktl, L. 1983. On the use of cluster analysis for partitioning and allocating computational objects in distributed computing systems. In: J.E. Gentleman (ed.), *Computer Science and Statistics: the Interface*, pps. 361-364. North Holland, Amsterdam.

Plastria, F. 1986. Two hierarchies associated with each clustering scheme. Patt. Recog. 19: 193-196.

Popma, J., L. Mucina, O. van Tongeren and E. van der Maarel, 1983. On the determination of optimal levels in phytosociological classification. Vegetatio 52: 65-76.

Rachman, M.I. and S.Ja. Koz'yakov. 1986. A statistical method for comparison of two structures and its biological application. Biom. J. 2: 183-195.

Rahman, N.A. 1962. On the sampling distribution of the studentized Penrose measure of distance Ann. Human Genet. 26: 97-106.

Ratkowsky, D. and G.N. Lance. 1978. A criterion for determining the number of groups in a classification. Austral Comput. J. 10: 115-117.

Rogers, C.C.G. 1978. The probability that 2 samples in the plane have disjoint convex hulls. J. Appl. Prob. 15: 790-802.

Rohlf, F.J. 1975. Generalization of the gap test for the detection of multivariate outliers. Biometrics 31: 93-101.

Rose, M.J. 1965. Classification of a set of elements. Comput. J. 7: 208-224.

Ross, D.R. 1979. TAXON *Users Manual*, ed. P3. CSIRO, Division Computing Research, Canberra, A.C.T.

Roubens, N. 1978. Pattern classification problems and fuzzy sets. Fuzzy sets and systems 1: 239-253.

Roubens, M. 1982. Fuzzy clustering algorithms and their cluster validity. Eur. J. Oper. Res. 10: 294-301.

Rousseau, P. 1978. Maximum likelihood clustering of binary data sets. Classif. Soc. Bull. 4.

Ruspini, E.H. 1982. A new approach to clustering. Inf. Control. 15: 22-32.

Sandland, R.L. and P.C. Young. 1979. Probabilistic tests and stopping rules associated with hierarchical classification techniques. Austral. J. Ecol. 4: 399-406.

Sankoff, D. and J.B. Kruskal. 1983. ''Time Warps, String Edits and Macromolecules: the theory and practice of sequence comparison''. Addision Wesley, London, pps. 382.

Sattath, S. and A. Tversky. 1977. Additive similarity trees. Psychometrika 42: 319-345.

Särndal, C.E.A. 1976. A Monte Carlo study of some asymmetric association measures. Brit. J. Math. Statist. Psychol. 29: 94-102.

Schaeben, H. 1984. A new cluster algorithm for orientation data. Math. Geol. 16: 139-153.

Scher, A., M. Shneier, and A. Rosenfeld. 1982. Clustering of collinear line segments. Patt. Recog. 15: 85-91.

Schueler, L. and H. Wolff. 1980. Automatic classification in the case of unknown number of clusters using global density estimates. Biom. J. 22: 745-754.

Schultz, J.V. and L.J. Hubert. 1973. Data analysis and the connectivity of random graphs. J. Math. Psychol. 10: 421-435.

Schultz, J.V. and L.J. Hubert. 1975. Empirical evalutation of an approximate result in random graph theory. Brit. J. Math. Statist. Psychol. 28: 103-111.

Sclove, S.L. 1977. Population mixture models and clustering algorithms. Commun. Statist.-Theor. Meth. A6: 417-434.

Scott, A.J. and M. Knott. 1976. An approximate test for use with AID. Appl. Statist. 25: 103-109.

Scott, D.W. and J.R. Thompson. 1983. Probability density estimates in higher dimensions. In: J.E. Gentleman (ed.), *Computer Science and Statistics, the Interface*, pps. 173-179. North Holland, Amsterdam.

Segen, J. and A.C. Sanderson. 1979. A minimal representation criterion for clustering. In. J.F. Gentleman (ed.), *Comput. Science and Statistics, the Interface*. pps. 332-334, North Holland, Amsterdam.

Selem, S.Z. and M.A. Ismail. 1984. Soft clustering of multidimensional data: a semi-fuzzy approach. Patt. Recog. 17: 559-568.

Selkow, S.M. 1974. Diagnostic keys as a representation for context in patterm recognition. IEEE Trans. Comput. C-23: 970-971.

Sen Gupta, A. 1982/83. Tests for simulataneously determining the number of clusters and their shape with multivariate data. Statist. Probab. Lett. 1: 46-50.

Shafer, E., R. Dubes and A.K. Jain. 1979. Single-link characteristics of a mode-seeking clustering algorithm. Patt. Recog. 11: 65-70.

Shanley, R.J. and M.A. Mahtab. 1976. Delineation and analysis of clusters in orientation data. Math. Geol. 8: 9-23.

Shapiro, L.G. and R.M. Haralick. 1979. Decomposition of two-dimensional shapes by graph-theoretic clustering. IEEE Trans. Patt. Anal. Mach. Intel. PAMI 1: 10-20.

Shepard, R.N. and P. Arabie. 1979. Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychol. Rev. 86: 87-123.

Siemiatychi, J. 1978. Mantel's space-time clustering statistic. I. Computing higher moments and a comparison of various data transforms. J. Statist. Comput. Simulation 7: 13-31.

Simon, J.C. and G. Guiho. 1972. On algorithms preserving neighbourhood to file and retrieve information in a memory. Intern. J. Comput. Inform. Sci. 1: 3-15.

Smith, S.P. and R. Dubes. 1980. Stability of a hierarchical clustering. Patt. Recog. 12: 177-187.

Smith, S.P. and A.K. Jain. 1984. Testing for uniformity in multidimensional data. IEEE Trans. Patt. Anal. Mach. Intel. PAMI-6: 73-81.

Smith, W. and J.F. Grassle. 1977. Sampling properties of a family of diversity measures. Biometrics 33: 282-292.

Sneath, P.H.A. 1966. A method for curve seeking from scattered points. Comput. J. 8: 383-391.

Sneath, P.H.A. 1979. BASIC program for a significance test for 2 clusters in Euclidean space as measured by their overlap. Comput. Geosci. 5: 143-155.

Sneath, P.H.A. 1980a. Some empirical tests for significance of clusters. In: E. Diday, L. Lebart, J.P. Pagès and R. Tomassone (eds.), *Data Analysis and Informatics*, pps. 491-508. North Holland, Amsterdam.

Sneath, P.H.A. 1980b. The probability that distinct clusters will be unrecognised in low dimensional ordinations. Classif. Soc. Bull. 4: 22-43.

SNEATH, P.H.A. 1985. DENBRAN: a BASIC program for a significance test for multivariate normality of clusters from branching points in dendrograms. Comput. Geosci. 11: 767-785.

SNEATH, P.H.A. 1986. Significance tests for multivariate normality of clusters from branching patterns of dendrograms. J. Math. Geol. 18: 3-32.

SONQUIST, J.A., E.L. BAKER and J.N. MORGAN. 1973. *Searching for Structure: An Approach to Analysis of Substantial Bodies of Micro-data and Documentation for a Computer Program*. Inst. Soc. Res., Univ. Michigan, Ann Arbor. 236 pps.

STEPP, R.E. and R.S. MICHALSKI. 1986. Conceptual clustering of structured objects: a goal orientated approach. Art. Intell. 28: 43-70.

STODDARD, A.M. 1979. Standardization of measures prior to cluster analysis, Biometrics 35: 765-773.

STRAUSS, R.E. 1982. Statistical significance of species clusters in association analysis. Ecology 64: 634-639.

SWITZER, P. 1968. Statistical techniques in pattern recognition and clustering. Proc. Amer. Statist Assoc. 40-47.

THURSTONE, L.L. 1945. A multiple group method for factoring the correlation matrix. Psyvchometrika 1: 73-78.

TOU, J.T. 1979. DYNOC - A dynamic optimal cluster-seeking technique. Internl. J. Comput. Inform. Sci. 8: 541-547.

TOUSSAINT, G.T. 1974. Some properties of Matusita's measure of affinity of several distributions. Ann. Inst. Statist. Math. 26: 389-394.

TRIVEDI, M.M. and J.C. BEZDEK. 1986. Low-level segmentation of aerial images with fuzzy clustering. IEEE Trans. Syst., Man Cybern. SMC-116: 589-598.

TSUKAMURA, M. 1976. Conditions for normal distribution of matching coefficients involved in a cluster in numerical classification. Japan. J. Microbiol. 20: 357-359.

UTTLEY, A.M. 1970. The INFORMON. J. theor. Biol. 27: 31-45.

VELASCO, F.R.D. 1980. A method of analysis of Gaussian-like clusters. Patt. Recog. 12: 381-393.

VERHELST, N.D., M.G.M. KOPPEN and E.P. VAN ESSEN. 1985. The exact distribution of and index of agreement between partitions. Brit. J. Math. Statist. Psychol. 38: 44-57.

VESELY, A. 1981. Logically oriented cluster analysis. Kybernetika 17: 82-92.

WACKER, A.G. 1972. The effect of subclass numbers on maximum likelihood gaussian classification. Proc. 8-th Remote Sensing Conf., East Lansing, pps. 851-859.

WAINER, H. and S. SCHACHT. 1978. Gapping. Physchometrika 43: 203-212.

WALLACE, C.S. and D.A. BOULTON. 1968. An information measure for classification. Comput. J. 11: 185-194.

WARNEKAR, C.S. and G. KRISHNA. 1979. An algorithm to detect linearly seperable clusters of binary variables. Patt. Recog. 11: 109-113.

WATANABE, S. 1969. *Knowing and Guessing*. J. Wiley, New York. pp 376-379.

WHITFIELD, J.W. 1953. The distribution of total rank values for one particular object in m rankings of n objects. Brit. J. Statist. Pshychol. 6: 35-40.

WILLIAMS, W.T. and J.S. BUNT. 1980. Studies on the analysis of data from Australian tidal forests (''Mangroves''.). II. The use of an asymmetric monothetic divisive classificatory program. Austral. J. Ecol. 5: 391-396.

WILLIAMS, W.T. and M.B. DALE. 1965. Fundamental problems in numerical taxonomy. Adv. Bot. Res. 2: 35-68.

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

WILLIAMS, W.T., J.M. LAMBERT and G.N. LANCE. 1966. Multivariate methods in plant ecology V. Similarity analyses and information analysis. J. Ecol. 54: 427-445.

WILLIAMS, W.T., G.N. LANCE L.J. WEBB, J.T. TRACEY and M.B. DALE. 1969. Studies in the numerical analysis of complex rain-forest communities III. The analysis of successional data. J. Ecol. 57: 515-535.

WILLIAMS, W.T. and J.G. TRACEY. 1984. Network analysis of north Queensalnd rainforests. Austral. J. Bot. 32: 109-116.

WINDHAM, M.P. 1985. Numerical classification of proximity data with assignment measures. J. Classif. 2: 157-172.

WISHART, D. 1969. Numerical classification method for deriving natural classes. Nature 22: 97-98.

WONG, M.A. 1984. Asymptotic properties of univariate sample K-means clusters. J. Classif. 1: 265-270.

WONG, A.K.C. and D.E. GHAHRAMAN. 1980. Random graphs: structural-contextual dichotomy. IEEE. Trans. Patt. Anal. Mach. Intel. PAMI-2: 1341-355.

WONG, A.K.C. and T.S. LIU. 1975. Typicality, diversity and feature pattern of an ensemble. IEEE. Trans. Comput. C-24: 158-181.

YAMAMOTO, S., K. USHIO, S. TAZAWA, H. IKEDA, F. TAMARI and N. HAMADA. 1977. Partitions of a query set into minimal number of subsets having the consecutive retrival property. J. Statist. Planning Infer. 1: 41-51.

YOLKINA, V.N. and N.G. ZAGORNIKO. 1978. Some classification algorithms developed at Novosibirsk. RAIRO Informatique/Computer science. 12: 37-46.