Methodology Review: Clustering Methods

Glenn W. Milligan and Martha C. Cooper
Ohio State University

A review of clustering methodology is presented, with emphasis on algorithm performance and the resulting implications for applied research. After an overview of the clustering literature, the clustering process is discussed within a seven-step framework. The four major types of clustering methods can be characterized as hierarchical, partitioning, overlapping, and ordination algorithms. The validation of such algorithms refers to the problem of determining the ability of the methods to recover cluster configurations which are known to exist in the data. Validation approaches include mathematical derivations, analyses of empirical datasets, and monte carlo simulation methods. Next, interpretation and inference procedures in cluster analysis are discussed. Inference procedures involve testing for significant cluster structure and the problem of determining the number of clusters in the data. The paper concludes with two sets of recommendations. One set deals with topics in clustering that would benefit from continued research into the methodology. The other set offers recommendations for applied analyses within the framework of the clustering process.

Classification is a basic human mental process. Relevant groupings can provide economy of memory, predictive power, or possible theoretical development. Much classificatory activity is carried out at a subjective level. However, with the advent of high-speed computational equipment, many disciplines have been involved in the development of automatic or objective algorithms for the generation of classifications. Before reviewing the literature, four issues peculiar to the area deserve discussion:

1. Many researchers are not aware of the immense amount of activity in the field of classification. Blashfield and Aldenderfer (1978) noted that the number of articles using clustering methodology grew from 25 in 1964 to 501 in 1976. Between 1958 and 1973, more than 1,600 scholarly articles on classification were published. A separate literature search indicated that in 1985 alone, 1,658 references were found on the topic.

   Despite the wide base of interest in the development and use of clustering methodology, the literature on the topic is remarkably segmented, with authors in one academic discipline often isolated from research in other fields. Blashfield (1976, 1980) and Blashfield and Aldenderfer (1978) have documented the lack of cross-reference between disciplines. It is not unusual to find identical or highly similar techniques being discovered in different fields and given different names.

   The richness of this literature base presents problems for the interested reader. Most discussions of clustering procedures are embedded in content material specific to a given discipline. A reader may have to deal with articles in soil science, quantitative biogeography,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 4, December 1987, pp. 329–354
0146-6216/87/040329-26$2.55

cladistics, inorganic molecular structures, market segmentation, or consistency theorems of mathematical statistics. The interdisciplinary nature of the sources makes it difficult for an applied researcher to become fully informed about developments in this methodological area. Nevertheless, it is not reasonable to ignore the contributions made in other fields.

2. Defining the term "cluster" presents a problem. Numerous definitions for the concept exist, and each can be valid within a particular application or framework. Some researchers have viewed clusters as mixtures of multivariate normal populations (Blashfield, 1976; Fleiss & Zubin, 1969; Wolfe, 1970). Each cluster can be viewed as a multivariate normal population. Because each cluster forms a different population, it is possible to conceptualize a "mixture" of populations present in the database available for sampling. This represents a logical extension of the use of the normal distribution in other cases. Such mixtures allow for clusters to overlap in the variable space.

   Milligan (1980, 1985) used truncated multivariate normal mixtures to ensure that clusters did not overlap. A truncated multivariate normal population would involve constraining all data points to fall within a specified interval about the mean for a given variable. For example, each point could be required to fall within ±2.0 standard deviations of the mean. This has the effect of eliminating the tails of the distribution. Assuming that the centroids for the populations are separated by a sufficient minimum distance, the populations will not overlap in the variable space. (The centroid is the location that corresponds to the means of the variables in the multivariate space.) The concept of distinct groups was incorporated by Cormack (1971) into a definition of natural clusters. The definition states that clusters should exhibit internal cohesion and external isolation. Several other authors have expressed this concept in similar terms. Sneath (1969) stated that "in a broad sense clusters are thought of as collections of points which are relatively close, but which are separated by empty regions of space from other clusters" (p. 260).

   However, the concept of natural clusters may be overly restrictive for some applications and inappropriate for others. If all elements within a given cluster are required to be highly similar to each other, then more elongated clusters are ruled out. When all data points are required to be highly similar, the result is a compact cluster approximating a sphere in the variable space. Thus, it may be desired to modify the concept of a compact cluster to allow for clusters that are elongated or continuously connected. Such elongated clusters could occur, for example, if the cluster follows a regression effect between variables. Similarly, requiring clusters to be disjoint might be an inappropriate model of a human population. A definition which allows for overlapping clusters is needed in such cases. Thus, when attempting a cluster analysis of an empirical dataset, researchers must address the issue of the definition of the concept of a cluster. This is essential because different clustering algorithms attempt to find different kinds of clusters.

3. Further difficulties are presented by the heuristic nature of the clustering methods themselves. Unlike regression, ANOVA, and even factor analysis, there is no single analysis technique based on some widely accepted statistical principle. The problem is compounded by the computational task involved. For the fairly small problem of dividing 25 elements into 5 nonoverlapping clusters, there are over 2.4×10^{15} different possible solutions (Anderberg, 1973). As such, each method must attempt to find the optimal clustering, using its own definition of cluster structure and optimality, without testing all possible partitions. There is no guarantee that a given algorithm will find the optimal partition in the data. To compound the problem, there are literally hundreds of clustering algorithms in existence. Because no single method is known to be optimal and so many methods are available for use, a literature has developed on the problem

of validating the accuracy of clustering algorithms.

4. Inference procedures have been developed for cluster analysis. For example, procedures exist to test whether significant cluster structure has been found in the data or whether there is a partition of random noise data containing no such structure. This is an important problem because virtually all clustering algorithms give solution partitions regardless of the presence or the absence of structure in the data. A different but related problem deals with the task of determining the number of clusters. This issue has been approached from many perspectives.

In large part, the remainder of this review addresses the issues of algorithm validation and the development of inference procedures. Before turning to the literature, a discussion of the clustering process and a general survey of the types of clustering algorithms are presented.

Steps in the Clustering Process

A seven-step structure is used to organize the clustering process. This structure is consistent with the discussions found in Anderberg (1973), Cormack (1971), Everitt (1980), and Lorr (1983). An applied article using clustering should contain information on the actions taken by the researcher for each of these steps. For the purposes of this paper, a clustering method will refer to the specific means by which entities are grouped together. Ward's minimum variance or *K*-means procedures are examples of clustering methods. The clustering process refers to the steps outlined in this section which represent the sequence necessary for a complete analysis. Implications of the decisions involved in each of these steps are discussed in later sections.

1. The entities to be clustered must be selected. The sample of elements should be chosen to be representative of the cluster structure in the population.
2. The variables to be used in the cluster analysis are selected. Again, the variables must contain sufficient information to permit the clustering of the objects.
3. The researcher must decide whether or not to standardize the data. If standardization is to be performed, then the researcher must select a procedure from several different approaches.
4. A similarity or dissimilarity measure must be selected. These measures reflect the degree of closeness or separation between objects. A dissimilarity measure, such as distance, assumes larger values as two objects become less similar. A similarity measure, such as correlation, assumes larger values as two objects become more similar.
5. A clustering method must be selected. The researcher's concept of what constitutes a cluster is important because different methods have been designed to find different types of cluster structures.
6. The number of clusters must be determined. This problem has received increased attention in the clustering literature during the last several years.
7. The last step in the clustering process is to interpret, test, and replicate the resulting cluster analysis. Interpretation of the clusters within the applied context requires the knowledge and expertise of the researcher's particular discipline. Testing involves the problem of determining whether there is a significant clustering or an arbitrary partition of random noise data. Finally, replication determines whether the resulting cluster structure can be replicated in other samples.

Although variations on this seven-phase process may be necessary to fit a particular application, this sequence represents the critical steps in a cluster analysis. The next section describes the various clustering methods which can be selected for use in step 5.

Types of Clustering Methods

With several hundred clustering methods in existence, some means of classification is needed to describe the techniques. Four major categories can be identified: hierarchical methods, partitioning (nonhierarchical) algorithms, overlapping clustering procedures, and ordination techniques.

Hierarchical Methods

Perhaps the most popular clustering algorithms have been the sequential agglomerative hierarchical methods. Such hierarchical methods begin with each entity considered as a separate cluster. At each successive level in the clustering, two of the clusters are merged. The clustering continues until only one cluster, containing the entire dataset, remains. The routine will generate a strictly nested "hierarchy" of n partitions, where n is the number of entities in the dataset. The partitions represent nonoverlapping clusters and have the property that once two elements become members of the same cluster, they are never again separated. The researcher has the option of using the entire hierarchy as the solution, or selecting a level representing the specific number of clusters of interest.

Different hierarchical methods are distinguished by the criterion for determining which two clusters to merge at each level. Lance and Williams (1967) demonstrated that many agglomerative hierarchical methods are variations of a common recurrence formula. Details concerning the recurrence formula can be found in Cormack (1971), Everitt (1980), Lorr (1983), and Milligan (1979). A more extensive and recent reference is Gordon (1987).

The use of agglomerative clustering accelerated in the 1960s as computers became available. Of particular interest to the psychological community is a series of articles published by McQuitty in *Educational and Psychological Measurement* (see McQuitty, 1987). Other important articles from this period would include the introduction of Ward's (1963) minimum variance method and Johnson's (1967) discussion of the complete- and single-link methods. Finally, D'Andrade (1978) introduced a routine based on the nonparametric U statistic.

At least one other major type of hierarchical clustering strategy exists. Divisive clustering methods follow a pattern that is the reverse of the agglomerative techniques. Divisive methods begin with all entities in one cluster and partition the data into two or more clusters. This process can be allowed to continue until n clusters are present which contain individual entities from the dataset. The Edwards and Cavalli-Sforza (1965) method attempts to find the division which minimizes the within-cluster error sum of squares. Unfortunately, divisive methods face problems of computational complexity which are not easily overcome (see Anderberg, 1973).

Partitioning Methods

Partitioning methods produce distinct nonoverlapping clusters. Often, the methods are known as nonhierarchical clustering procedures because only a single data partition is produced (Anderberg, 1973; Sneath & Sokal, 1973; Späth, 1980). The techniques range in complexity from Hartigan's (1975) very simple leader algorithms to rather intricate iterative reallocation methods, such as ISODATA (Ball & Hall, 1965), Friedman and Rubin's (1967) method, and NORMIX (Wolfe, 1970).

Partitioning methods can be distinguished by five characteristics (Blashfield, 1977a). The first characteristic involves the selection of the initial starting partition or "seed points." Some K-means methods use randomly selected data elements as starting partitions, while others allow the user to specify starting seeds (Anderberg, 1973; Jancey, 1966). Finally, Wolfe's (1970) NORMIX routine uses Ward's hierarchical method to start the algorithm.

The second and third characteristics deal with the type of cluster assignment pass made through the data and the statistical criterion used to assign the points to the clusters. Some K-means algorithms make a single pass, assigning each point in turn to the nearest cluster centroid, while others make multiple passes and update the centroids after each point assignment. Similarly, the statistical criteria range from a simple distance measure between a point and a cluster centroid, to an attempt to optimize rather complex matrix criteria borrowed from multivariate normal distribution theory (see Marriott, 1971; Scott & Symons, 1971).

The final two features involve whether a fixed or variable number of clusters will be formed, and the eventual treatment of outliers in the solution.

Most methods require the user to specify the number of clusters, and outliers are forced to join one of the clusters present in the solution. Only a few methods, such as ISODATA, allow for a variable number of clusters in the solution, or for a residual pool of unassigned points.

Overlapping Methods

Compared with the first two categories of clustering methods, there is a much smaller number of algorithms which allow for overlapping clusters. At times, these are called clumping or clique formation methods. An early algorithm which produced overlapping structures was published by Needham (1967). Many authors working in this area have used graph theoretic concepts to help develop overlapping methods. Examples include Ling (1973), Ozawa (1985), and the more axiomatic approach of Jardine and Sibson (1971). Methods first introduced in the psychological literature include the ADCLUS procedure developed by Shepard and Arabie (1979), as well as the Hubert (1974) and Peay (1975) methods. Finally, Corter and Tversky (1986) have introduced a technique for identifying overlapping clusters by a graphical representation of extended trees.

Ordination Methods

Although most social scientists are unfamiliar with the term ordination, psychologists have been primarily responsible for developments in this area. The term is more commonly used in the biological sciences and statistics (see Cormack, 1971). Ordination techniques attempt to provide some type of dimensional representation, usually based on fewer variables than in the original dataset. Thus, techniques such as factor analysis (*R*- and *Q*-mode) and metric and nonmetric multidimensional scaling fall into this category. Extensive discussions of factor-analytic procedures as applied to classification can be found in Cattell (1952) and Tryon and Bailey (1970). One feature of ordination methods is that only a spatial representation of the entities in the dataset is produced. The actual determination of cluster membership is left to the researcher's subjective judgment.

Validation Techniques

Because all clustering methods are heuristic in nature, the critical issue of recovery performance must be addressed. That is, it is necessary to verify that a given method can recover the true cluster structure in a dataset. Unless a method can be shown to reliably recover known configurations in error-free data, it will not be very useful in applied analyses. Furthermore, it is desirable to determine the sensitivity of such methods to different forms of error in the data or to errors in judgment made during the clustering process. A literature addressing these concerns exists. Generally, three strategies have been used (Dubes & Jain, 1979, 1980): (1) mathematical or theoretical derivations, (2) analysis of empirical datasets, or (3) monte carlo simulation. A discussion of the advantages and limitations of each approach follows.

Mathematical Derivation

An analytical or theoretical derivation would be the method of choice. Unfortunately, the overwhelming complexity of the process has limited advances in this area. A few articles have appeared which use graph theory (see Matula, 1977). Others have taken a statistical approach (e.g., Bock, 1985; Hartigan, 1985). Still others have addressed the problem from a geometric perspective (Gower, 1967; Milligan, 1979). However, progress has been slow, and derivational results have often had limited value for applied analyses; thus, work in this area has had little impact on the ''end user'' applications of cluster analysis.

A different problem with the derivational approach occurs when theoretical recommendations conflict or run counter to applied experience with the methods. For example, Jardine and Sibson (1971) developed an elegant axiomatic system for defining an acceptable clustering method. As it turned out, only the single-link hierarchical method could satisfy the requirements of the system. This result generated a heated controversy because experience with the single-link method indicated that it was one of the poorest performing algorithms (Williams, Lance, Dale, & Clifford, 1971).

Similarly, Fisher and Van Ness (1971) proposed a set of nine admissibility criteria to evaluate clustering methods. Again, the single-link method appeared to be quite satisfactory, whereas nonhierarchical *K*-means methods rated rather poorly. However, Wong (1982) has made successful use of the *K*-means approach in a hybrid clustering scheme. Milligan (1980) found the *K*-means methods to give exceptionally good recovery when well-chosen starting centroids were used. Hence, theoretical findings are available which conflict with results obtained from other approaches. Discrepancies of this sort led Dubes and Jain (1980) to conclude that the derivational approach was less useful as a validation strategy in the clustering context.

Analysis of Empirical Datasets

By far the most common validation strategy has been the application of a clustering method to an empirical dataset. Typical examples of such analyses are found in Goldstein and Linden (1969) and Harrigan (1985). Often, only one clustering method is tested on the data, hence no comparative information is offered to the reader. In fact, the most common way to introduce a new clustering method into the literature is to test it on an applied dataset (see, e.g., Johnson, 1967).

However, this approach has a serious and potentially fatal weakness. If the clustering algorithm finds the subjective clustering that the researcher suspects exists in the data, then some evidence indicates that the algorithm may be able to find the correct structure; however, the evidence is based on a sample of size one. More seriously, if the clustering results deviate from prior expectations, then the discrepancy is difficult to explain. The researcher's subjective clustering may have been incorrect, and the structure found by the algorithm may be the correct one. On the other hand, the algorithm may have missed the correct cluster structure, or there may be no structure in the data at all. Given the methodological weakness of this approach, results from such studies should be treated with some skepticism.

Simulation Analysis

A simulation or monte carlo study typically requires three major steps or phases. First, artificial datasets with known cluster structure are selected or generated. Second, these constructed datasets are analyzed by the various clustering methods or procedures of interest in the study. Finally, the level of agreement between the known cluster structure and the structure found by the clustering procedures is determined through the use of one or more recovery indices. The results reported from such simulation studies are usually based on summary statistics or inference procedures computed from the recovery indices.

In the first phase of the analysis, artificial data are generated which contain a known cluster structure. Several methods exist for the construction process, including use of a pencil and graph paper. However, the most commonly employed method is to write a computer program which generates the datasets. Before such a program is written, the researcher must decide on some definition or conceptualization of the clusters. For example, the clusters might be viewed as samples from a mixture of multivariate normal populations in a specified variable space. During the preparation of the data generation program, routines are designed for specifying various characteristics of the clusters. These include decisions concerning the number of clusters, the number of dimensions, the number of elements per cluster, the centroids for the clusters, and the variance-covariance matrices for the populations from which the clusters are sampled.

These characteristics have a direct impact on the nature of the resulting clusters. For example, if cluster centroids are required to be widely spaced in the multivariate space, then generally nonoverlapping clusters will be obtained. Otherwise, an overlapping structure can be generated. Once the features of the clusters have been determined, the simulation program uses a random number generator to sample a set of points that will comprise each cluster. Finally, the researcher may introduce various forms of error or noise into the data. For example, outliers and random noise dimensions could be added to study their effect on various clustering procedures. Examples of cluster generation routines that have seen repeated use include Blashfield (1976) and Milligan (1985).

The second phase is to analyze the constructed datasets using the clustering methods or procedures of interest. Often, the computer code for the clustering procedures is modified to eliminate extraneous output and to store the information about the resulting cluster solution. Despite the program modification, this phase is not overly complicated. The analysis of the constructed data is comparable to analysis of data provided by the same procedures in applied research.

In the final phase, two partition sets are obtained for each constructed dataset. The first partition is that which was used to define the clusters in the data generation program. This partition is the true cluster structure of the data. The second partition set is the one obtained from the clustering procedure, and corresponds to the partition which would have been used if this had been an analysis of an applied (empirical) dataset. The most convenient way to present the results of the simulation analysis has been to compute some measure of agreement between the true partition structure and the obtained clusters. These measures have been called recovery or consensus indices.

One measure which has seen active use is the Rand index (see Hubert & Arabie, 1985). Both the numerator and denominator of the index reflect frequency counts. The numerator involves taking each pair of elements and determining whether the classification of the pair is consistent between the known and obtained clusterings. That is, the points must be treated in the same manner in both partition sets. If the pair of points is in the same cluster in both the known and obtained clusterings, then the frequency count for the numerator is increased by 1.0. Similarly, if the points are in different clusters in both partitions, the numerator count is increased by 1.0. The denominator is the total number of possible pairwise comparisons and equals $n \times (n - 1)/2$, where n is the number of elements. If the obtained clustering exactly matches that of the known partitioning, then the numerator and denominator are equal and the index value is 1.0. If any discrepancies are found, then the numerator is less than the denominator and the index value is less than 1.0. The smaller the index value, the greater the inconsistency between the known and obtained clusterings.

The Rand index is but one of several recovery measures that have been proposed. The process of selecting a recovery measure can be complex—for example, the original version of the Rand index is no longer recommended for use in simulation studies—and the literature has begun to address the topic (e.g., Day, 1986; Hubert & Arabie, 1985; Milligan & Cooper, 1986).

A distinct advantage of the simulation approach is that there is no doubt as to the true cluster structure. It is possible to examine recovery across hundreds of datasets, thus avoiding conclusions based on a single dataset. Furthermore, the process circumvents the mathematical complexities found in attempting a derivational study. The main disadvantage is the limited generalizability to data distributions and structures which were not considered in the simulation experiments. This last problem can be troublesome for an applied researcher attempting to select a method for an empirical analysis. However, the simulation literature has contributed some of the clearest evidence about method performance and will have a dominant impact on the review of validation results in the next section.

Validation Results

The following review of the validation results is organized around the four categories of clustering methods. The majority of the findings relate directly to clustering method performance. In general, results will be based on the simulation literature. As discussed previously, the selection of the clustering method for an applied analysis represents the fifth step in the clustering process. Decision errors made in the other steps also can cause reduced recovery performance. Thus, an additional section presents the results obtained when specification errors are made in the other steps of the clustering process. These include the impact of outliers when selecting data elements (step 1), selection of variables (step 2), standardization (step 3), and dissimilarity measures (step 4).

Hierarchical Methods

Agglomerative hierarchical clustering procedures have been the most frequently examined. At least 11 systematic studies have been conducted; results from these experiments are summarized in Table 1. The most commonly tested algorithms have been the single link, complete link, group average, and Ward's method. In addition, recovery information concerning Lance and Williams' (1967) beta-flexible method has been included in the table. Other hierarchical routines have been examined, such as the centroid and median methods; however, the performance of these methods has not been particularly noteworthy and they have not been listed. Additional smaller-scale or more specialized studies (Blashfield & Morey, 1980; Cunningham & Ogilvie, 1972; D'Andrade, 1978; Gross, 1972; Morey, Blashfield, & Skinner, 1983; Rand, 1971) were also omitted.

When examining Table 1, and subsequently Table 2, it is important to note that the criterion index used in the experiments may differ from study to study. Thus, direct numerical comparisons of criterion values can be made only within a study and not across reports. However, all criteria have the property that perfect recovery of the true cluster structure would generate values of 1.0. As such, the closer the average or median criterion value is to 1.0, the better the recovery performance of the algorithm. The values presented in the tables have been selected as typical representations of the respective experiments, or reflect summary statistics for an entire study.

One of the earliest systematic studies was conducted by Baker (1974). Baker examined three hierarchical structures consisting of chained, binary, and random-link trees. A total of 100 datasets were generated for each tree type and error condition. Findings presented in Table 1 are for the binary tree and are typical of the overall results. Only the single- and complete-link methods were considered. Baker found that the performance of the single-link method was severely impaired by the presence of increased error perturbation of the interpoint distances. The complete-link method exhibited less sensitivity to this factor. Even in error-free data, the complete-link method usually gave better recovery performance than the single-link procedure.

Kuiper and Fisher (1975) conducted a more extensive study which examined the impact of a variety of design factors on cluster recovery. In general, bivariate normal mixtures were used to generate the artificial datasets. A total of 30 datasets were created for each experimental condition. Selected results from the study appear in Table 1. The first two lines for the Kuiper and Fisher study in the table correspond to the case when the clusters were of equal size. In this case, Ward's method produced the best recovery of the underlying structure. However, when unequal size clusters were present in the data, the complete-link and group-average methods gave superior recovery. Except for this last condition, Ward's method appeared to be the best clustering procedure for recovering clusters from bivariate normal mixtures. Finally, the single-link method produced recovery which was significantly inferior to any other method.

The next study listed in Table 1 was performed by Blashfield (1976). Unlike many researchers, Blashfield used nonzero covariances and more realistic principal component structures to construct 50 datasets. Multivariate normal mixtures with complex covariance patterns were used to generate 2 to 6 clusters embedded in a space of 3 to 22 dimensions. Cluster sizes varied from 5 to 40. These characteristics were selected randomly for each dataset. In retrospect, it is unfortunate that these features were not systematically controlled in an experimental context. Recovery was averaged over datasets with differing numbers of clusters and with clusters of unequal sizes. Nevertheless, Blashfield found that Ward's method gave significantly better

Table 1
Validation Results for Hierarchical Clustering Methods

	Method				
Study	Single Link	Complete Link	Group Average	Ward's	Beta Flexible
Baker (1974)					
Low Error	.605	.968			
Medium Error	.298	.766			
High Error	.079	.347			
Kuiper & Fisher (1975)					
Medium Size	.579	.742	.710	.767	
Five Clusters	.444	.690	.630	.707	
Unequal Sizes	.663	.705	.702	.689	
Blashfield (1976)					
	.06	.42	.17	.77	
Mojena (1977)					
	.369	.637	.596	.840	
Mezzich (1978)					
Correlation	.625	.973			
Euclidean	.648	.943			
Edelbrock (1979)					
Correlation	.90	.80	.96		
Euclidean	.62	.63	.70	.88	
Milligan & Isaac (1980)					
	.30	.64	.70	.57	
Bayne et al. (1980)					
Configuration 1	.53	.68	.66	.70	
Configuration 2	.55	.76	.75	.76	
Edelbrock & McLaughlin (1980)					
Correlation	.858	.813	.880		
Euclidean	.690	.780	.858	.873	
Milligan (1980)					
Zero Error	.974	.995	.998	.987	.997
Low Error	.902	.970	.997	.989	.994
High Error	.777	.880	.948	.940	.945
Scheibler & Schneider (1985)					
Correlation	.43	.49	.81	.78	.73
Euclidean	.04	.38	.16	.79	.77

recovery performance than any other procedure. The complete-link method was the next best, with the group-average method a distant third. Finally, as found with the previous studies, the single-link method gave the poorest recovery performance. (It should be noted that the data possessed overlapping clusters; Milligan, 1981b, found that cluster overlap favored Ward's technique over other methods.)

Edelbrock (1979) conducted a simulation study using 10 of the 50 datasets generated by Blashfield (1976). Rather than insisting on recovery of the exact number of the clusters, Edelbrock argued in favor of reduced coverage. Solutions in the hierarchy that contained more clusters than actually existed in the data were examined. The results presented in Table 1 were obtained at the 90% coverage level. Recovery did improve when lower coverage was allowed. Edelbrock found a significant advantage for the use of the correlation similarity measure. The best recovery was obtained from the group-average method using correlation. Ward's method was not tested with the correlation measure, but its performance using Euclidean distance was quite good.

Mojena (1977) generated 12 datasets consisting of clusters based on mixtures from multivariate gamma populations. The univariate gamma probability function represents a fairly rich family of distributions that includes, as a special case, the chi-square family. As such, gamma populations are not generally symmetric; as the mean increases, the variance increases. For the multivariate case, Mojena set all of the covariances to 0. Clusters were of equal size, and consisted of 30 points each. Mojena systematically varied the degree of cluster overlap in the experiment. The results in Table 1 represent mean recovery values for each method. The overall ranking and performance of the methods is quite consistent with that of Blashfield (1976). Ward's method gave significantly better recovery than any other procedure. The single-link method performed significantly worse. Mojena did find that as the level of overlap increased, the impact on Ward's method was less severe than the impact on the group-average algorithm.

Mezzich's (1978) study has received fairly extensive coverage. However, only two artificial datasets were examined. The results in Table 1 show that the complete-link procedure was much more effective at recovering the cluster structure than the single-link method. Unlike Edelbrock (1979), the study found little impact from the use of differing dissimilarity measures, and the effect was not consistent from method to method.

Milligan and Isaac (1980) generated clusters that corresponded to ultrametric tree structures. Ultrametric distances are more restrictive than Euclidean distances. In an ultrametric space, the distances between any three points must form an equilateral or isosceles triangle where the base is shorter than the two equal sides. No such restriction applies to Euclidean space (see Milligan, 1979). Various nonoverlapping configurations and error levels were included in the experiment. Typical results from the study are presented in Table 1. The authors found that the group-average method gave the best overall recovery. The complete-link procedure was second best, with Ward's method a distant third in recovery performance. The performance of the single-link method was found to be seriously impaired by the presence of error in the data. This characteristic caused the method to give the poorest overall recovery. Finally, as would be expected, recovery declined as the separation between clusters decreased for all methods.

Bayne, Beauchamp, Begovich, and Kane (1980) used a variety of parameterizations of two bivariate normal populations. Each parameterization was based on 200 constructed datasets. The results presented in Table 1 represent typical realizations from their experiment after converting their criterion to a percentage-correct value. Overall, the results showed that Ward's method gave the best performance, with the group-average and complete-link algorithms placing a close second. The single-link method performed much worse than all other methods. The authors found that the performance of the group-average method declined rapidly as the separation between populations decreased. Ward's method was less affected by this factor.

Edelbrock and McLaughlin (1980) reported a study based on the same principles and logic as the Edelbrock (1979) experiment. For validation purposes, a total of 20 datasets from Blashfield (1976) and 12 datasets from Mojena (1977) were used. As can be seen in Table 1, the results were basically the same as those found by Edelbrock (1979). Edelbrock and McLaughlin also examined additional similarity measures which involved one- and two-way intraclass correlation coefficients. The one-way intraclass correlation was found to provide enhanced recovery performance for the group-average method.

In an extensive validation experiment, Milligan (1980) examined 11 hierarchical clustering procedures, including the five listed in Table 1. A total of 108 error-free datasets were created which consisted of truncated multivariate normal mixtures.

The generation process ensured that the clusters were nonoverlapping in the variable space. Two additional error conditions were created which involved perturbing the interpoint distances at low and high levels. The results in Table 1 indicate that the group-average and beta-flexible methods gave the best recovery, with Ward's method a close second. None of the other hierarchical algorithms examined in the experiment produced superior recovery rates. The error perturbation process showed a greater impact on the complete-link procedure, and a very marked impact on the single-link method.

The last and most recent study was conducted by Scheibler and Schneider (1985). The authors generated 200 constrained normal mixtures. Again, all five methods listed in Table 1 were tested. Ward's procedure and the beta-flexible methods performed well using either correlation or Euclidean distance measures. However, the group-average method gave excellent recovery when using the correlation index, but performed badly with Euclidean distance. This result is similar to that found in the Edelbrock (1979) study, but is less similar to the results of Edelbrock and McLaughlin (1980). It is not consistent with the results of Milligan (1980, 1981b). Finally, the complete-link and single-link methods (especially the latter) gave poor recovery of the underlying cluster structure.

Several conclusions can be drawn from the series of experiments summarized in Table 1. Ward's method tended to perform well in the cases where it was tested. Often, it gave the best cluster recovery. The performance of the group-average method was more erratic. At times, it produced the best recovery of cluster structure. However, its performance was not good on other occasions. The reasons for the discrepant recovery behavior have not yet been identified. On the other hand, the beta-flexible method performed well in the few studies where it has been included. The enhanced performance pattern of the beta-flexible method was confirmed in a recent study which systematically varied the beta parameter across a wide range of values (Milligan, 1987a). The complete-link algorithm has occasionally performed better than the group-average method, but it is usually inferior to Ward's procedure. The single-link method, while theoretically attractive, has been repeatedly shown to give poor cluster recovery and to be seriously affected by the presence of even small levels of error in the data.

Partitioning Methods

A fairly substantial validation literature exists for nonhierarchical procedures. Five major systematic studies have been conducted; their results are compared in Table 2. The first study was conducted by Blashfield (1977a) and was based on 20 multivariate normal mixture datasets generated in an earlier study (Blashfield, 1976). The best recovery was obtained from three methods which gave similar median recovery values. These were the CLUSTAN *K*-means procedure with random starting seeds, and the two methods that used the $|\mathbb{W}|$ criterion. (For purposes of this paper, $\mathbb{W}$, $\mathbb{B}$, and $\mathbb{T}$ represent the within-cluster, between-cluster, and total sum of squares and cross-products matrices, respectively.) It is difficult to explain the reduced performance level (.643) of the CLUSTAN *K*-means procedure when the starting centroids were obtained from Ward's method. Given the results for Ward's method from Blashfield (1976), these centroids should have been fairly accurate.

In Mezzich's (1978) study, all but one partitioning procedure gave recovery values which did not differ much from method to method, as seen in Table 2. These recovery values were equivalent to the results for the complete-link hierarchical method in Table 1 for Mezzich. The best partitioning technique was a *K*-means procedure using Euclidean distances. The only procedure which did not produce equivalent recovery was Wolfe's (1970) NORMIX method. It performed rather poorly despite the fact that centroids from Ward's method were used as the starting seeds for the algorithm.

Bayne et al. (1980) also considered four partitioning methods, as listed in Table 2, in addition to the hierarchical methods listed in Table 1. The authors found that the convergent *K*-means method and the two versions of the Friedman and Rubin algorithm gave the best recovery of all methods

Table 2
Validation Results for Nonhierarchical Clustering Methods

Clustering Method	Criterion
Blashfield (1977)	
Forgy K-means	.585
Convergent K-means	.638
CLUSTAN K-means	.706 (.643)
Friedman-Rubin Trace W	.545
Friedman-Rubin |W|	.705
MIKCA Trace W	.560
MIKCA |W|	.699
Mezzich (1978)	
Convergent K-means: Correlation	.955
Convergent K-means: Euclidean Distances	.989
Ball-Hall ISODATA	.977
Friedman-Rubin |W|	.966
Wolfe NORMIX	.443
Bayne et al. (1980)	
Convergent K-means	.83
Friedman-Rubin Trace W	.82
Friedman-Rubin |W|	.82
Wolfe NORMIX	.70
Milligan (1980): Low Error Condition	
MacQueen K-means	.884 (.934)
Forgy K-means	.909 (.996)
Jancey K-means	.926 (.993)
Convergent K-means	.901 (.996)
Scheibler & Schneider (1985)	
CLUSTAN K-means	.67 (.78)
Späth K-means	.55 (.77)

Note. Parenthetical entries indicate recovery performance for K-means methods when starting centroids were obtained from Ward's or group average hierarchical clustering procedures. Otherwise, randomly selected data were used as seed points.

tested. The performance of the *K*-means and |W| criterion is consistent with the studies by Blashfield (1977a) and Mezzich (1978). However, the enhanced performance of the Trace W criterion is not in accord with the Blashfield study. Finally, Wolfe's (1970) NORMIX method gave rather poor recovery, which is similar to Mezzich's result.

Milligan (1980) conducted a study which included four partitioning *K*-means methods. The results indicated that the methods were sensitive to the nature of the starting partition, but in a manner different than that found by Blashfield (1977a). When randomly selected data elements were used as starting seeds, the four *K*-means procedures gave reduced recovery values. The results for the random starting condition are given as the first column of values for Milligan (1980) in Table 2. When centroids obtained from the group-average hierarchical method were used as starting seeds, the procedures gave excellent recovery. The parenthetical entries in Table 2 for Milligan give the recovery values when the starting centroids were based on the group-average method. The four K-means methods seemed to be equivalent except for MacQueen's procedure, which tended to give lower recovery values than the other three methods.

The recent study by Scheibler and Schneider (1985) included two different versions of the K-means algorithms. The results indicated that both clustering methods gave better recovery when centroids from Ward's procedure were used to specify initial seed points. This result is consistent with Milligan (1980) but not with Blashfield (1977a). Finally, the CLUSTAN K-means method gave better recovery than Späth's (1980) technique. (It should be noted that Späth's procedure uses medians rather than centroids to locate the clusters.)

The success of the K-means procedures with improved starting seeds found in some studies has led several authors to propose hybrid algorithms for applied clustering (Milligan & Sokol, 1980; Punj & Stewart, 1983). These procedures use the centroids obtained from a hierarchical method to start a K-means algorithm. A different hybrid model was suggested by Wong and Lane (1983). The first stage involves a Kth nearest-neighbor density estimation process followed by a hierarchical clustering of the nearest-neighbor distance matrix. Example analyses in Wong and Lane suggest good recovery performance.

In summary, the convergent K-means method tended to give the best recovery of cluster structure. This result was obtained despite the theoretical findings of Fisher and Van Ness (1971), which indicated that the K-means procedures were fairly unattractive methods. Furthermore, method sophistication or complexity may have little impact on the quality of the obtained solution. The K-means procedures are fairly direct and simple, whereas the methods using the $|W|$ criterion are rather complex. Yet the methods tended to give equivalent recovery. The results for the Trace W criterion were inconsistent. Bayne et al. found it to be among the best methods tested, whereas Blashfield found the criterion to give the lowest recovery values. The reduced performance level is consistent with less systematic studies of clustering criteria (Friedman & Rubin, 1967; Scott & Symons, 1971). Finally, Wolfe's NORMIX method performed poorly in all cases where it was tested.

Overlapping Methods

It appears that no systematic validation study of overlapping clustering methods has been conducted. Jardine and Sibson (1971) presented a single clustering of 23 Indian caste groups to demonstrate their B_k overlapping clustering method. Jardine and Sibson apparently felt that their axiomatic derivation provided sufficient justification for the algorithm. Hubert (1974) used a dataset based on 13 diagnostic psychiatric categories to study his two-diameter clustering procedure. Peay (1975) used two datasets to demonstrate his method, one relating to the study of marriages in eight ethnic groups, and the other an artificially constructed dataset of six elements. Shepard and Arabie (1979) studied ADCLUS with five datasets, three of which dealt with similarities or confusions in letters or numbers based on set sizes of 10, 16, and 26. The remaining two datasets studied a communication network of 14 industrial workers and results based on sorting names of 20 anatomical terms. Finally, Ozawa (1985) used a constructed dataset of eight points. Included for comparison against his own algorithm was Jardine and Sibson's B_k method. Both methods found the same correct structure.

Several conclusions can be drawn from the literature. First, the information that is available on the performance of overlapping methods is based on the analysis of empirical datasets. As such, recovery performance is difficult to judge. Second, the example datasets possessed rather small sample sizes; the average sample size in the studies was only 14.4. This may not be representative of the more typical applications of clustering. Finally, without comparative information, it is uncertain which method may have the more desirable recovery characteristics.

Ordination Methods

Social scientists have been using or exploring the properties of ordination methods for several decades (Blashfield, 1980). Among the early advocates of the use of inverted or *Q*-factor analysis for clustering applications were Cattell (1952) and Tryon (see Tryon & Bailey, 1970). Certainly, if the clusters exist in the reduced factor space, and if these clusters correspond to the type of simple structure that most modern orthogonal factor rotation programs seek, then a *Q*-analysis should allow the user to identify the correct cluster structure. However, if any of the prerequisite conditions cannot be satisfied, then recovery may be marginal at best. For example, if the clusters are found in or defined by an oblique factor space, then an orthogonal varimax rotation may make it difficult for the user to determine correct cluster membership. A discussion of other methodological difficulties follows.

First, assume that clusters exist in the space defined by the original dataset. Sneath (1980) has shown that there is a high probability that a researcher will conclude that a subset of points comprises one cluster, when in fact the points comprise two or more clusters. The reduction in dimensionality produced by the *Q*-analysis impairs the user's ability to detect clusters that existed in the space defined by the original variables.

Fleiss, Lawlor, Platman, and Fieve (1971) reached a parallel conclusion in their study of inverted factor analysis. These authors felt that some indication of distinct grouping should be present in the original data before a *Q*-analysis is attempted. If evidence for clustering existed, such as multimodal or nonsymmetric variables, then Fleiss et al. further concluded that methods other than inverted factor analysis might do a better job at finding the clusters. Blashfield and Morey (1980) confirmed this expectation in a study of simulated MMPI profiles. They found that both the group average and Ward's hierarchical methods gave better recovery and were less problematic in application than *Q*-factor analysis. Mezzich (1978) reached similar conclusions when using simulated psychiatric profiles.

A different issue in the use of ordination methods is the subjective nature of the cluster identification task. Different researchers may interpret the output from the same ordination analysis as forming different groups. Mezzich (1978) studied the interrater reliability for determining cluster membership with nonmetric multidimensional scaling. Reliability was found to average about .77 using Cramer's statistic. Mezzich did not measure interrater reliability when clusters were derived from an inverted factor analysis. It would seem reasonable to assume that moderate to low reliability values also would be found for this technique.

Given these and other logical difficulties, many researchers have expressed reservations about the use of ordination methods as clustering procedures. Lorr (1983) noted that the appropriateness of *Q*-analysis has long been in dispute. Even one of the early advocates of the method, Cattell (1978), has stressed the fact that *Q*-analysis is not a procedure for finding types (clusters), but a technique for finding dimensions. Similar comments would hold for other ordination methods, such as multidimensional scaling.

Ordination methods have been proposed as a strategy for data preparation prior to the application of a clustering procedure. This factoring or scaling would then be part of the third step in the clustering process. Some authors would suggest a dimensional analysis if a large number of variables has been collected in a study. Rather than conducting a *Q*-analysis on the entities, a regular factor analysis would be performed on the variables. However, if the clusters were defined (or existed) in the original variable space, then an ordination method would serve to distort or hide the true structure as shown in Sneath's (1980) study. On the other hand, if the clusters existed in the reduced variable space, then an ordination method should be useful for detecting the true clustering.

Kaufman (1985) reported the results of a simulation study where the clusters were first defined and generated in the reduced variable space. Five strategies for preprocessing the data were studied. Four strategies were based on principal components analysis, taking into account whether all components were used or just those with eigenvalues above 1.0, and whether the scores were weighted according to the corresponding eigenvalues. A fifth condition involved using the original standardized variables directly with the clustering method. All processed datasets were clustered with Ward's hierarchical method. The results indicated that weighted principal components analysis gave the best recovery of the correct cluster structure.

This result is logical, given the way in which the data were constructed. However, it is interesting to note that simple standardization performed nearly as well as the more complex analysis. Thus, the assumption that principal components processing is a necessity was not confirmed, even when the data were constructed to be most favorable for such an analysis. Kaufman failed to include a condition where the unstandardized data were analyzed directly. Such a condition would have addressed the question of whether standardization itself was necessary.

Other Validation Results

Validation results exist for a variety of other aspects or steps in the clustering process. These include the effect of outliers and the corresponding issue of coverage, improper selection of a dissimilarity measure, inclusion of random noise dimensions in the variable set, standardization of variables, and the impact of overlapping data structures on nonoverlapping clustering methods.

Outliers. Objects which are not members of any cluster in the dataset can be considered outliers to all clusters. Of course, such elements could be intermediates between clusters and not outliers to the entire data mass. Milligan (1980) examined the effect of adding 20% or 40% additional data points as outliers. As would be expected, the presence of outliers served to confuse or mask the assignment of data elements which did belong to a cluster. Edelbrock (1979) argued that, from a psychological perspective, it is not necessary for all persons in a dataset to be classified in order to obtain useful partitions. Edelbrock found that reduced coverage resulted in improved recovery performance with a set of hierarchical methods. Similar results were obtained in the study by Scheibler and Schneider (1985), which included nonhierarchical routines. However, both studies were based on the use of the kappa statistic, and there is a possibility that the results on reduced coverage were confounded with characteristics of this recovery measure (Milligan, 1987a).

Similarity measures. In terms of the clustering process, it is important to note that the use of the wrong dissimilarity measure for a dataset might lead to reduced recovery of the cluster structure present in the data. Hence, some care should be taken when selecting a dissimilarity measure. For example, if the clusters are embedded in a Euclidean space, then a Euclidean distance dissimilarity measure would be appropriate. A three-dimensional Euclidean space corresponds to that normally encountered by people in their day-to-day activities. Blashfield (1977b) warned that several formulas for Euclidean distance exist in the clustering literature. For purposes of the present discussion, the following formulation will be used:

$$d_{ij} = \left[\sum w_k (x_{ik} - x_{jk})^2 \right]^{1/2} . \qquad (1)$$

In Equation 1, d_{ij} is the ''straight-line'' distance between objects i and j. The summation is computed over all variables in the dataset; k is the index value for the summation operator. The values x_{ik} and x_{jk} represent the variable values for objects i and j, respectively, on variable k. The values w_k are weights applied to the squared difference on each variable and are usually set to 1.0. A priori weights determined by the researcher can also be used. Other weighting schemes exist; a particularly effective procedure will be discussed below.

Euclidean distance is a special case of the more general Minkowski family of metric distances. Another member of the family is the so-called city block or Manhattan distance. Rather than measuring straight-line distance, the city-block measure would determine the distance by finding the shortest path in a grid system, similar to that taken by an automobile through city streets.

A different measure of similarity which has been popular in the social sciences is the Pearson correlation computed between two objects. The Pearson correlation is the cosine of the angle between two vectors representing the standardized scores of the objects. The correlation measure has the advantage (or disadvantage, depending on the context) of eliminating the effects of differing means and variances between variables when computing the similarity measure (see Cronbach & Gleser, 1953; Skinner, 1978). Neither Euclidean distance nor the city-block measure eliminates these differences. Many other measures have been proposed; see the discussions in Anderberg (1973), Cormack (1971), Everitt (1980), and Lorr (1983) for a more complete introduction to this topic.

In the validation literature, a number of researchers have considered the consequences of the use of alternative or incorrect dissimilarity measures. The research indicates that differing dissimilarity measures can change the extent of cluster recovery. However, Punj and Stewart (1983) have argued that errors in the choice of a dissimilarity measure do not seem to be as serious as other decision errors in the clustering process.

Irrelevant variables. A few experiments have been conducted to determine the impact of adding variables to the dataset which are irrelevant to the cluster structure. These variables can be viewed as random noise dimensions. Milligan (1980) investigated the impact of adding one or two irrelevant dimensions to the set of variables which define the true clustering. The results indicated that the addition of even one irrelevant variable seriously reduced the extent of cluster recovery. Thus, it seems ill-advised to indiscriminately include variables in a dataset for an applied cluster analysis.

The use of all available data will likely obscure any clustering present in a subset of the variables. Similar masking effects were obtained by Fowlkes in two unpublished studies (see De Soete, DeSarbo, & Carroll, 1985). These results led De Soete et al. to develop an optimal weighting scheme for variables in a hierarchical cluster analysis. That is, values other than 1.0 are assigned to the weights w_k in Equation 1 when computing the Euclidean distance between points. The weights provide an optimal fit between the resulting distances and an ultrametric tree structure. The logic behind the optimal fit is that most hierarchical clustering methods force the ultrametric structure on the clustering solution. Results obtained by Milligan (1987b) indicated that the De Soete et al. algorithm is quite effective in assigning near-zero weights to irrelevant variables. Recovery of true cluster structure was enhanced in all cases examined, and the impact of the masking effect was greatly reduced.

Standardization. A researcher must decide whether to standardize the variables in a dataset. Edelbrock (1979) found a slight advantage for standardized data when all elements were required to be clustered. When coverage was reduced, standardization produced no improvement in cluster recovery. However, Milligan (1980) found that standardization can lead to a limited reduction in recovery performance when the clusters exist in the unstandardized space. In a more recent study, Milligan and Cooper (in press) conducted a large-scale simulation study of eight standardization procedures. Results of the experiment indicated that standardization procedures based on division by the range of the variable were consistently more effective than any other approach, including the traditional *z*-score procedure. The Milligan and Cooper paper includes a fairly complete review of the issues on the topic and references to the literature.

Overlapping structures. Some researchers have considered the problem of cluster recovery by hierarchical and partitioning methods when the data consist of overlapping clusters. Effectively, there is a mismatch between the clustering algorithm selected and the structure of the data. Of course, in an applied analysis, a researcher may not be aware that an overlapping structure is present in the data. Mojena (1977) conducted a small experiment where the overlap between clusters was systematically varied. As the extent of overlap increased, the proportion of correct cluster assignments decreased. Similarly, half of the datasets generated by Blashfield (1976) possessed overlapping characteristics (see Milligan & Isaac, 1980). Unfortunately, the results were not broken down according to this characteristic.

This led Milligan (1981b) to conduct an experiment which directly compared performance on overlapping and disjoint structures. Ward's hierarchical method was found to give the best performance when overlap was allowed to exist in the data. The group-average method tended to perform better with well-separated structures. These results are consistent with those of Bayne et al. (1980). Scheibler and Schneider (1985) reported that they could not confirm the finding. However, Scheibler and Schneider did not examine overlapping structures in their experiment.

It appears that no studies have considered the reverse situation. A study of the performance of overlapping methods on data consisting of disjoint clusters would be of interest. An ''optimal'' clustering algorithm would produce an overlapping or a disjoint clustering solution, as appropriate for the structure in the data.

Inference, Replication, and Interpretation Procedures

Several streams of research can be identified which have addressed issues relating to the last two steps in the clustering process. The following sections deal with the problem of determining the number of clusters (step 6), hypothesis testing in a clustering context, replication analysis, and aids to interpretation (the last three comprise step 7).

Determining the Number of Clusters

Most clustering procedures require the user to specify or to determine the number of clusters in the final solution. A substantial segment of the literature which addresses this decision task is referenced in Milligan and Cooper (1985). These authors conducted a simulation study of the performance of 30 decision rules. Although the experiment used only hierarchical clustering methods, all decision rules are adaptable to nonhierarchical algorithms. Only well-separated, distinct clusters were present in the structure of the 108 test datasets.

Given such well-defined clustering, it would be expected that the performance of any rule for determining the number of clusters would be fairly good. Milligan and Cooper did find a set of reasonably effective rules. This set included the techniques developed by Baker and Hubert (1975), Calinski and Harabasz (1974), Duda and Hart (1973), Beale (1969), and the cubic clustering criterion used in SAS (Sarle, 1983). The set of techniques is fairly heterogeneous and includes measures adapted from classical parametric and nonparametric procedures. The Calinski and Harabasz, Duda and Hart, and Beale indices use various forms of sum of squares within or between clusters, or both.

The Milligan and Cooper study also revealed that a number of methods did not work well. Among methods which performed poorly were Trace W, criteria based on |W|, and a generalized distance measure. Unfortunately, the Trace W (error sum of squares) criterion has been the most frequently suggested method. Further, most criteria borrowed from traditional multivariate normality approaches, such as Wolfe's (1970) likelihood ratio test, gave at best only mediocre correct detection rates. Everitt (1981) found that Wolfe's (1970) likelihood ratio test performed well only for large sample sizes.

In a more limited study, Begovich and Kane (1982) confirmed that a criterion based on |W| performed rather poorly. In contrast to Milligan and Cooper, the authors found that the Calinski and Harabasz procedure did not work well when the clusters possessed unequal covariance matrices. A different strategy, called simulation cluster analysis (SCA), gave the best performance in their study. SCA is similar to parallel analysis in factor analysis, and involves creating a series of datasets by adding random error to the original data. Each error-perturbed dataset is subjected to a cluster analysis and the results from the simulations are combined to provide a likelihood estimate of the number of groups in the data.

Milligan and Cooper did not consider techniques for determining the number of clusters which required the user to make subjective judgments (see, e.g., Gower, 1975). Similarly, techniques which are dependent on the use of a specific clustering method were not studied. Hence, potentially useful methods such as that of Wong and Schaak (1982) have not been independently validated.

Hypothesis Testing

Despite Romesburg's (1984) assertion to the contrary, hypothesis testing in a cluster-analytic situation can certainly be performed. Studies by Bock (1985), Hartigan (1977, 1978, 1985), Lee (1979), and Sneath (1977) indicate that many researchers have been addressing the distributional problems in cluster analysis. Most testing procedures have been developed to determine whether significant cluster structure exists in the partitions found by a clustering method. The null hypothesis typically specifies that the data consist of a random pattern of points with no distinct clustering present. For example, the hypothesis may specify that the data were sampled from a single multivariate normal population, or the data correspond to what would be expected from a uniform distribution contained in a hypercube.

In general, the testing procedures can be divided into two approaches (Sneath, 1969). In the first approach, it is assumed that an external criterion variable is available to validate the clustering results. The second approach, using an internal criterion, uses information which is contained within the cluster analysis. Before considering the two approaches, comments are in order regarding the naive application of standard hypothesis testing procedures in a cluster-analytic framework.

After clustering results have been obtained, it is tempting to conduct a discriminant analysis, or some other standard testing procedure such as ANOVA or MANOVA. Examples of such applications can be found as early as Turner (1969) and as recently as Andes (1986). The logic is that if significant differences exist between the groups found in the cluster analysis, then significant clustering must be present in the data. The variables used in the cluster analysis are employed as the dependent variables in the tests, and the cluster partitions define the treatment groups for the testing procedure.

It is widely recognized by clustering methodologists that this strategy is invalid, but this has not been stressed in the literature (see Dubes & Jain, 1979; Milligan & Mahajan, 1980; Morey et al., 1983). Such procedures will almost always return significant test results, even for random noise data that contain no cluster structure. The reason for the bias when using discriminant analysis is that the groups were not defined a priori. More seriously, the variables tested were the same ones used to define the groups in the first place. Similarly, in the case of MANOVA, there is systematic violation of the assumption of random assignment of observations to the different groups.

The cluster analysis process does not haphazardly assign the elements to groups, but assigns them to maximize the similarity between observations within the groups. Most clustering algorithms will find homogeneous groups of points in random noise data, even though such partitions do not represent any significant clustering property. Fortunately, the testing procedures which follow are valid and avoid the problems in the use of traditional testing methods.

External criterion analysis. An external criterion represents information which is external to the cluster analysis. That is, the information in the external criterion was not used at any other point in the clustering process. This information could be in the form of one or more variables which can help validate the grouping, or in terms of a partition of the elements into groups. The partition must be either specified a priori or obtained from a clustering of a separate dataset. When the external criterion is in the form of one or more variables, it is permissible to perform a standard parametric analysis such as ANOVA or MANOVA to validate the clustering results. Unfortunately, many applied researchers find it difficult to omit such variables from the cluster analysis if they believe that the variables contain information regarding the true cluster structure. Hence, few studies have used the external criterion approach.

If the researcher has a criterion which represents an independent partitioning of the data, then a procedure developed by Hubert and Baker (1977) can be used to test the validity of the clustering results. It is important to note that the Hubert and Baker procedure cannot be used to test the similarity of two clusterings of the same dataset.

Internal criterion analysis. An internal criterion measure uses information obtained from within the clustering process. Internal criterion measures typically reflect the goodness of fit between the input dissimilarity matrix and the resulting clustering. Hence, the strength of the clustering with respect to the input data is measured. Two specific examples of such measures are the Baker and Hubert (1975) gamma index and the point-biserial correlation measure (Milligan, 1980).

The gamma index is the ratio of two frequency counts. The numerator is the difference between the counts for consistent and inconsistent pairings of distances. The denominator is simply the sum of the number of pairings of both kinds. A consistent pairing occurs when a within-cluster distance is found to be smaller than a between-cluster distance. An inconsistent pairing is recorded when a within-cluster distance is larger than a distance between two points not in the same cluster. Thus, the gamma index has a value of 1.0 when the clustering solution is perfectly consistent with respect to between- and within-cluster distances.

The point-biserial measure is a Pearson correlation between a variable coded 0 or 1 and a variable corresponding to the input distances. For a given input distance between two points, the binary variable is assigned a value of 0 if the resulting clustering solution placed the two points in the same cluster. Otherwise, a 1 is recorded, denoting a between-cluster distance. Larger values of the point-biserial measure would indicate greater agreement between the clustering of the data and the input distances.

Reviews of internal criteria can be found in Cormack (1971), Jardine and Sibson (1971), and Rohlf (1974). A comparative study of 30 internal criteria was conducted by Milligan (1981a). Milligan found a set of highly effective internal criteria which included the gamma index and the point-biserial measure. It was found that Trace W and criteria adapted from the multivariate normality literature (such as those based on functions of W and $|W|$) performed poorly. These results do not give much support to the testing procedures proposed by Arnold (1979) using $\log(|T|/|W|)$ as the test statistic.

Once an effective internal criterion has been selected, it can be used as a test statistic. The statistic is used to test an alternative hypothesis that specifies that some form of significant cluster structure exists in the obtained data partitions. The main difficulty in conducting the test is to determine an appropriate sampling distribution for the statistic. One approach has been to use permutation or monte carlo procedures to generate an approximate sampling distribution. Milligan and Sokol (1980) developed such a test based on the point-biserial criterion. Other authors have adopted this approach, including Begovich and Kane (1982) and Good (1982).

Replication Analysis

A different approach, based on the logic of cross-validation as used in regression analysis, was proposed by McIntyre and Blashfield (1980):

1. Two samples are obtained for clustering purposes. This can be accomplished by randomly dividing one larger dataset into two samples.
2. The first sample is cluster-analyzed, and the centroids for the clusters are computed. (This step assumes that a determination of the number of clusters was made.)
3. The distance (or other dissimilarity measure) is computed between each element in the second sample and each of the centroids determined from the clustering of the first sample.
4. Each element in the second sample is assigned to the nearest cluster centroid from the first sample. This last assignment activity produces a clustering of the second sample based on the characteristics of the first sample.
5. The second sample is directly cluster-analyzed using its own data. There are now two clusterings of the second sample for comparison purposes.
6. Some measure of agreement is computed between the two partitions of the same data. The consistency between the original cluster so-

lution and the cross-validated cluster assignments indicates the stability of the solution.

McIntyre and Blashfield suggested the use of the kappa statistic as a measure of consistency. The authors provided some initial information on the numerical values of kappa that would occur in situations where the agreement between the partition sets was high. However, Hubert and Arabie's (1985) corrected Rand index seems to exhibit better properties for the comparison of partitions (Milligan, 1987a). Unfortunately, no typical estimates for the Hubert and Arabie index have been published for the cross-validation application.

A different strategy involves comparisons of clusterings from fairly different sources. However, these comparisons usually are based on a single dataset. For example, it is possible to compare partitionings from different clustering methods. A level obtained from a hierarchical clustering method could be selected, and the clusters could be compared to those found by a nonhierarchical procedure. If the cluster structure remained fairly consistent across different clustering methods, it would seem reasonable to conclude that the structure is strong and not an artifact of any given method. In fact, comparisons are not restricted to those involving the same number of clusters in the two partition sets, although in most situations the equal-number case would be the most interesting to an applied researcher. Finally, comparisons based on the use of different similarity or dissimilarity measures are also possible.

These more general comparisons introduce a level of confounding in the results if a successful replication does not occur. A failure to replicate may be due to a lack of structure in the data, or to differences in the types of structures that differing clustering methods impose on the resulting solutions. More general comparisons could provide useful information if the context of the clustering problem suggested that the selected comparison was logical and meaningful.

The use of a single sample in the comparisons comes at a very heavy cost to the researcher, in that the ability to generalize the clustering results to other datasets is lost. The subtle shift from two- to one-sample comparisons rather dramatically changes the analysis from one involving external validity to one giving results only on internal consistency.

Aids for Interpretation

A very important characteristic of any applied cluster analysis is the interpretability of the resulting partitions. Because interpretation is dependent on the underlying context of the problem, the responsibility for this part of the analysis falls upon the user. However, clustering methodologists have developed a variety of techniques to help with the interpretation task.

Logically, descriptive statistics for each cluster should be computed. Any understanding of the nature of the variables employed in the analysis can be used for interpretive purposes. Note that any preprocessing of the input variables, such as standardization or principal components, would make this activity more difficult. A different approach suggested by Anderberg (1973) involves permuting the dissimilarity matrix according to the groups specified by a cluster analysis. That is, the rows and columns are reordered in such a manner as to place points within the same cluster in consecutive order. This should form a block diagonal matrix. The within-block values would correspond to the within-cluster distances, and the remaining entries would be between-cluster distances. If distinct clusters exist in the data and the clustering method detected them, then the within-block distances should be small compared to those outside the block.

In fact, most interpretive methods are graphical in nature. For example, Kruskal and Landwehr (1983) developed a method for presenting the results of a hierarchical cluster analysis. Their technique, called ''icicle'' plots, is an improvement on the more standard dendrogram or skyline plots produced by clustering packages. More elaborate approaches were proposed by Kleiner and Hartigan (1981) based on natural-appearing ''trees'' and ''castles.'' Besides showing the clustering of the data points, the trees presented by these authors are able to encode additional information. This is accomplished by varying the thickness of the branches, the angle between branches, the length of the branch or stem, and the order of the elements as determined by the branch (left vs. right). However, as the comments and rejoinder which immediately follow the article indicate, there is no consistent agreement as to the most effective way of displaying clustering results. Finally, a different graphical approach, based on hierarchical trees but allowing for overlapping structures, was introduced by Corter and Tversky (1986).

There have been several attempts to develop mathematical approaches which measure the stability or replicability of the generated cluster solution. One such approach is the cluster validity profiles introduced by Bailey and Dubes (1982). Validity profiles are obtained for each cluster in the solution and indicate the relative compactness or isolation of the cluster. The profiles are scaled in probability units and are derived from the mathematical field of graph theory (also see Matula, 1977). The profiles are compared with the best compactness and isolation values that would be expected in a randomly chosen graph. The profiles help reject spurious clusters and appear to identify valid ones.

Discussion

Recommendations for Research in Clustering

Clearly, much remains to be done to more completely understand the characteristics and limitations of existing clustering methods, in addition to dealing with the introduction of new procedures. Research based on the following recommendations would provide results of direct benefit to researchers undertaking applied analyses.

1. More information is needed concerning the sensitivity of clustering procedures to differing data characteristics. Such an understanding could eventually result in guidelines which indicate the most appropriate clustering procedure to use, given the properties of the input dataset. This would include all components in the clustering process, such as the selection of the data elements, the similarity measure, and the clustering method. For example, the presence of outliers may affect the decision made at each step in the clustering process. Remarkably little research has been conducted on this issue, but it is a rather complex problem. A good review of the difficulties involved can be found in Soon (in press).
2. Little is known about the performance of the clustering routines that yield overlapping clusters. A careful examination of this issue is needed. It would be useful to determine whether any of the existing methods can reliably discriminate between disjoint structures and those that require an overlapping representation. Similarly, the degree of overlap that can be tolerated before algorithm performance degrades would be helpful information. Finally, the impact of various sources of error on cluster recovery for these methods is unknown.
3. More information is needed on the reliability of subjective decisions that must be made in conjunction with some clustering procedures. In particular, the determination of cluster membership is often left up to the researcher's judgment when ordination methods are used. If the assignment process is reliable, is it valid? Similarly, the impact of the various graphical procedures used in clustering is not well understood. Issues concerning which graph is best for displaying cluster membership, inter-cluster relationships, and other features have not been addressed in a thorough manner.
4. When new derivational or simulation experiments are conducted, one or more of the methods which have been widely used in past research should be included. This provides a basis for comparison and linkage with the previous literature. In particular, when a new clustering method is introduced into the literature, comparative performance information against existing procedures is essential for evaluating the usefulness of the contribution.

Recommendations for Applied Analyses

The clustering process is a complex sequence of tasks that must be carefully executed to obtain a proper clustering of the user's data. Furthermore, the literature in the area of classification is quite diverse, with contributions coming from all fields of science. Certainly, much progress has been made since Everitt (1979) published a paper on unresolved problems in cluster analysis. Among the concerns listed by Everitt were the problem of determining the number of clusters, selection of clustering method, graphical techniques, development of computer packages, more efficient algorithms, and the application of the jackknife (a permutation-like estimation procedure). Work in all of these areas is continuing. Using the results that are currently available, some guidelines can be offered to the researcher interested in performing a cluster analysis. These recommendations will be presented within the context of the seven-step clustering process. Although the overview is necessarily somewhat cursory, the critical decisions have been included.

1. The sample of the entities to be clustered should be representative of the population. The sample could be systematic in nature and not random. If it is suspected that a particular cluster consists of a fairly small proportion of the population, then this cluster might be over-represented in the sample in order to facilitate cluster detection and to obtain more stable estimates of cluster centroids. The user might consider the inclusion of persons who serve as markers or ''ideal types,'' that is, persons known or believed to be members of different clusters. Finally, it seems reasonable to delete outlier observations because these may degrade algorithm performance.
2. Care must be exercised in the selection of the variables used in the cluster analysis. Some applied researchers seem to believe that they should use as many variables as are available. However, selecting the correct variables is a critical part of the analysis. The addition of only one or two irrelevant variables can have a serious effect on cluster recovery. Thus, the inclusion of irrelevant variables should be avoided. The researcher might consider providing a justification for the inclusion of each variable in terms of how it should discriminate among clusters. The use of the optimal variable weighting scheme of De Soete et al. (1985) can be used to improve cluster recovery in those cases where it is uncertain which variables contribute to the clustering in the data. Finally, the practice of using principal components to preprocess the data may result in undetected clusters and interpretation problems with the reduced variable space.
3. The issue of variable standardization must be addressed. The routine application of standardization in all analyses is not necessarily appropriate, especially when the variables have similar means and variances. If sizable differences in means or variances do exist and these differences are not related to the clustering in the data, then standardization would seem to be necessary. Alternative standardization procedures which are not based on the well-known *z* score appear to be more effective (Milligan & Cooper, in press). In particular, standardization based on division by range of the variable may improve recovery when standardization is needed.
4. The researcher will need to select a similarity or dissimilarity measure. The measure should reflect those characteristics that are suspected to define the clusters believed to be present in the data. At times, a tailor-made index may be needed and this is permissible. In order for an applied user to understand the properties of various dissimilarity measures, a few typical profiles can be constructed. The dissimilarity measures can be computed between profiles, and the resulting values can be studied to determine whether the relevant characteristics are being reflected in the values. The present authors have found this to be an effective exercise.
5. A clustering method must be selected. Three aspects must be considered here. First, the clustering method should be one designed to recover the cluster types suspected to be present in the data. A mismatch between the cluster type being sought by the clustering algorithm

and the cluster types that actually exist in the data may result in reduced or distorted recovery. Second, the method should be effective at recovering the specified structure, even when error is present in the data. For hierarchical methods, this would tend to rule out the use of the single and complete methods in favor of Ward's method or the group-average technique. In the case of partitioning methods such as *K*-means, the use of fairly accurate starting seeds seems to be important.

Finally, access to computer software to run the selected method is crucial. More often than not, the actual clustering technique selected by an applied user is determined by this last factor, and not by the first two criteria. For example, D'Andrade's (1978) hierarchical method would be an attractive alternative to the single- and complete-link methods when the user possesses ordinal data. However, software to run D'Andrade's routine has not been made available. Fortunately, the major statistical computer packages have been improving their clustering software over time, and this is becoming less of a problem. The Statistical Analysis System (SAS Institute, 1985) is one such package that has greatly enhanced its clustering options over the past decade. Continued improvements are expected and desirable.

6. The number of clusters will need to be determined or specified for most clustering methods. Milligan and Cooper (1985) provided a useful summary and bibliography of most of the techniques that have been proposed to address this decision task. Again, SAS has been implementing many of the rules which have been found to be reasonably effective.
7. The last step involves the interpretation, testing, and replication of the clustering results. The introduction of improved graphical techniques for interpretation purposes has certainly been a welcome development in the clustering literature. Access to these procedures is still somewhat limited, as is access to the testing routines that will allow a user to test the hypothesis of significant clustering in the data.

Unfortunately, limited access exacerbates a longstanding problem in the practice of applied clustering. Often, an applied researcher brings to a cluster analysis the assumption that clusters actually do exist in the data. The idea of testing this assumption escapes some researchers. These same researchers would not assume, say in the context of a laboratory experiment, that the groups were significantly different as a result of the experimental manipulations. Rather, the researcher would test for this result. The same logic is valid in the clustering context. Finally, the two-sample cross-validation process proposed by McIntyre and Blashfield (1980) is an excellent strategy for establishing the generalizability of a cluster analysis.

References

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

Andes, N. (1986, June). *Validation of cluster solutions using discriminant analysis and bootstrap techniques*. Paper presented at the meeting of the Classification Society of North America, Columbus OH.

Arnold, S. J. (1979). A test for clusters. *Journal of Marketing Research, 19*, 545–551.

Bailey, T. A., & Dubes, R. (1982). Clustering validity profiles. *Pattern Recognition, 15*, 61–83.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques. Case I: Sensitivity to data errors. *Journal of the American Statistical Association, 69*, 440–445.

Baker, F. B., & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association, 70*, 31–38.

Ball, G. H., & Hall, D. J. (1965). *ISODATA, a novel method of data analysis and pattern classification*. Menlo Park CA: Stanford Research Institute. (NTIS No. AD 699616)

Bayne, C. K., Beauchamp, J. J., Begovich, C. L., & Kane, V. E. (1980). Monte carlo comparisons of selected clustering procedures. *Pattern Recognition, 12*, 51–62.

Beale, E. M. L. (1969). *Cluster analysis*. London: Scientific Control Systems.

Begovich, C. L., & Kane, V. E. (1982). Estimating the number of groups and group membership using simulation cluster analysis. *Pattern Recognition, 15*, 335–342.

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

Blashfield, R. K. (1977a). *A consumer report on cluster analysis software: (3) Iterative partitioning methods* (NSF grant DCR 74-20007). State College PA: Pennsylvania State University, Department of Psychology.

Blashfield, R. K. (1977b). The equivalence of three statistical packages for performing hierarchical cluster analysis. *Psychometrika, 42*, 429–431.

Blashfield, R. K. (1980). The growth of cluster analysis: Tryon, Ward, and Johnson. *Multivariate Behavioral Research, 15*, 439–458.

Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature of cluster analysis. *Multivariate Behavioral Research, 13*, 271–295.

Blashfield, R. K., & Morey, L. C. (1980). A comparison of four clustering methods using MMPI monte carlo data. *Applied Psychological Measurement, 4*, 57–64.

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification, 2*, 77–108.

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1–27.

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

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum Press.

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

Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika, 51*, 429–451.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing the similarity between profiles. *Psychological Bulletin, 50*, 456–473.

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

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

Day, W. H. E. (Ed.). (1986). Consensus classifications [Special issue]. *Journal of Classification, 3*(2).

De Soete, G., DeSarbo, W. S., & Carroll, J. D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least squares approach. *Journal of Classification, 2*, 173–192.

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition, 11*, 235–254.

Dubes, R., & Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. *Advances in Computers, 19*, 113–228.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Edelbrock, C. (1979). Comparing the accuracy of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research, 14*, 367–384.

Edelbrock, C., & McLaughlin, B. (1980). Hierarchical cluster analysis using intraclass correlations: A mixture model study. *Multivariate Behavioral Research, 15*, 299–318.

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

Everitt, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics, 35*, 169–181.

Everitt, B. S. (1980). *Cluster analysis* (2nd ed.). London: Heinemann.

Everitt, B. S. (1981). A monte carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research, 16*, 171–180.

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

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

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

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

Goldstein, S. G., & Linden, J. D. (1969). A comparison of multivariate grouping techniques commonly used with profile data. *Multivariate Behavioral Research, 4*, 103–114.

Good, I. J. (1982). An index of separateness of clusters and a permutation test for its statistical significance. *Journal of Statistical Computing and Simulation, 15*, 81–84.

Gordon, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A, 150*, 119–137.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics, 23*, 623–628.

Gower, J. C. (1975). Goodness-of-fit criteria for classification and other patterned structures. In G. Estabrook (Ed.), *Proceedings of the 8th International Conference on Numerical Taxonomy*. San Francisco: Freeman.

Gross, A. L. (1972). A monte carlo study of the accuracy of a hierarchical grouping procedure. *Multivariate Behavioral Research, 7*, 379–389.

Harrigan, K. R. (1985). An application of clustering for strategic group analysis. *Strategic Management Journal, 6*, 55–73.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hartigan, J. A. (1977). Distribution problems in clustering. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 45–71). New York: Academic Press.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics, 6,* 117–131.

Hartigan, J. A. (1985). Statistical theory in clustering. *Journal of Classification, 2,* 63–76.

Hubert, L. J. (1974). Some applications of graph theory to clustering. *Psychometrika, 39,* 283–309.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Hubert, L. J., & Baker, F. B. (1977). The comparison and fitting of given classification schemes. *Journal of Mathematical Psychology, 16,* 233–253.

Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany, 14,* 127–130.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. New York: Wiley.

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

Kaufman, R. L. (1985). Issues in multivariate cluster analysis: Some simulation results. *Sociological Methods and Research, 13,* 467–486.

Kleiner, B., & Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles (with comments and rejoinder). *Journal of the American Statistical Association, 76,* 260–276.

Kruskal, J. B., & Landwehr, J. M. (1983). Icicle plots: Better displays for hierarchical clustering. *The American Statistician, 37,* 162–168.

Kuiper, F. K., & Fisher, L. (1975). A monte carlo comparison of six clustering procedures. *Biometrics, 31,* 777–783.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: I. Hierarchical systems. *Computer Journal, 9,* 373–380.

Lee, K. L. (1979). Multivariate tests for clusters. *Journal of the American Statistical Association, 74,* 708–714.

Ling, R. F. (1973). A probability theory of cluster analysis. *Journal of the American Statistical Association, 68,* 159–164.

Lorr, M. (1983). *Cluster analysis for the social sciences*. San Francisco: Jossey-Bass.

Marriott, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics, 27,* 501–514.

Matula, D. W. (1977). Graph theoretic techniques for cluster analysis. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 95–129). New York: Academic Press.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 15,* 225–238.

McQuitty, L. L. (1987). *Pattern-analytic clustering*. New York: University Press of America.

Mezzich, J. (1978). Evaluating clustering methods for psychiatric diagnosis. *Biological Psychiatry, 13,* 265–346.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika, 44,* 343–346.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325–342.

Milligan, G. W. (1981a). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46,* 187–199.

Milligan, G. W. (1981b). A review of monte carlo tests of cluster analysis. *Multivariate Behavioral Research, 16,* 379–407.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika, 50,* 123–127.

Milligan, G. W. (1987a). *A study of the beta-flexible clustering method* (WPS 87-61). Columbus OH: Ohio State University, Faculty of Management Sciences.

Milligan, G. W. (1987b). *A validation study of a variable weighting algorithm* (WPS 87-111). Columbus OH: Ohio State University, Faculty of Management Sciences.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159–179.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21,* 441–458.

Milligan, G. W., & Cooper, M. C. (in press). A study of standardization of variables in cluster analysis. *Journal of Classification.*

Milligan, G. W., & Isaac, P. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition, 12,* 41–50.

Milligan, G. W., & Mahajan, V. (1980). A note on procedures for testing the quality of a clustering of a set of objects. *Decision Sciences, 11,* 669–677.

Milligan, G. W., & Sokol, L. M. (1980). A two-stage clustering algorithm with robust recovery characteristics. *Educational and Psychological Measurement, 40,* 755–759.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20,* 359–363.

Morey, L. C., Blashfield, R. K., & Skinner, H. A. (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research, 18,* 309–329.

Needham, R. M. (1967). Automatic classification in linguistics. *The Statistician, 17,* 45–54.

Ozawa, K. (1985). A stratificational overlapping cluster scheme. *Pattern Recognition, 18,* 279–286.

Peay, E. R. (1975). Nonmetric grouping: Clusters and cliques. *Psychometrika, 40,* 297–313.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*, 134–148.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Rohlf, F. J. (1974). Methods of comparing classifications. *Annual Review of Ecology and Systematics, 5*, 101–113.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont CA: Lifetime Learning Publications.

Sarle, W. S. (1983). *Cubic clustering criterion* (Tech. Rep. A-108). Cary NC: SAS Institute.

SAS Institute (1985). *SAS user's guide: Statistics, version 5 edition*. Cary NC: Author.

Scheibler, D., & Schneider, W. (1985). Monte carlo tests of the accuracy of cluster analysis algorithms—A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research, 20*, 283–304.

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics, 27*, 387–397.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87–123.

Skinner, H. A. (1978). Differentiating the contribution of elevation, scatter, and shape in profile similarity. *Educational and Psychological Measurement, 38*, 297–308.

Sneath, P. H. A. (1969). Evaluation of clustering methods. In A. J. Cole (Ed.), *Numerical taxonomy* (pp. 257–271). New York: Academic Press.

Sneath, P. H. A. (1977). A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap. *Mathematical Geology, 9*, 123–143.

Sneath, P. H. A. (1980). The risk of not recognizing from ordinations that clusters are distinct. *Classification Society Bulletin, 4*, 22–43.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco: Freeman.

Soon, S. C. (in press). On detection of extreme data points in cluster analysis. (Doctoral dissertation, Ohio State University, 1988.) *Dissertation Abstracts International*.

Späth, H. (1980). *Cluster analysis algorithms*. New York: Wiley.

Tryon, R. C., & Bailey, D. C. (1970). *Cluster analysis*. New York: McGraw-Hill.

Turner, M. E. (1969). Credibility and cluster. *Annals of the New York Academy of Sciences, 161*, 680–688.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Williams, W. T., Lance, G. N., Dale, M. B., & Clifford, H. T. (1971). Controversy concerning the criteria for taxonometric strategies. *Computer Journal, 14*, 162–165.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329–350.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association, 77*, 841–847.

Wong, M. A., & Lane, T. (1983). A *k*th nearest neighbor clustering procedure. *Journal of the Royal Statistical Society, Series B, 45*, 362–368.

Wong, M. A., & Schaak, C. (1982). Using the *k*th nearest neighbor clustering procedure to determine the number of subpopulations. *Proceedings of the Statistical Computing Section, American Statistical Association*, 40–48.

Author's Address

Send requests for further information to Glenn W. Milligan, Faculty of Management Sciences, 301 Hagerty Hall, Ohio State University, Columbus OH 43210, U.S.A.

Reprints

Reprints of this article may be obtained *prepaid* for $2.50 (U.S. delivery) or $3.00 (outside U.S.; payment in U.S. funds drawn on a U.S. bank) from Applied Psychological Measurement, N657 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.