Multivariate Behavioral Research, 1985, 20, 283–304

Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms: A Comparison of Hierarchical and Nonhierarchical Methods

Dieter Scheibler and Wolfgang Schneider
University of Heidelberg

Nine hierarchical and four nonhierarchical clustering algorithms were compared on their ability to resolve 200 multivariate normal mixtures. The effects of coverage, similarity measures, and cluster overlap were studied by including different levels of coverage for the hierarchical algorithms, Euclidean distances and Pearson correlation coefficients, and truncated multivariate normal mixtures in the analysis. The results confirmed the findings of previous Monte Carlo studies on clustering procedures in that accuracy was inversely related to coverage, and that algorithms using correlation as the similarity measure were significantly more accurate than those using Euclidean distances. No evidence was found for the assumption that the positive effects of the use of correlation coefficients are confined to unconstrained mixture models.

The generic term cluster analysis describes a large family of statistical classification procedures. Since the early 1960s, when high-speed computers made the use of this procedure relatively easy, more than 100 different clustering algorithms have been developed. Milligan (1981) conducted a computerized search of the literature in 1976 and showed that new or considerably revised algorithms were appearing at a rate of about one per month. Only a few authors (e.g., D'Andrade, 1978; von Eye & Wirsing, 1978, 1980) however, have tried to compare their new procedures with already existing clustering algorithms.

In view of the diversity of available algorithms, the potential consumer of cluster analysis faces several problems. First, because of the development of different terminologies in different fields of application, several—in some cases up to seven—labels are often used for the same clustering algorithm (Blashfield & Aldenderfer, 1978). More important, few guidelines are available for choosing a clustering procedure for research applications. This problem is especially perplexing since different algorithms are likely to produce different solutions when applied to the same data set (Bartko, Strauss, & Carpenter, 1971). Furthermore, there is no guarantee that any of the available clustering algorithms will recover the true cluster structure either under error-free or error-perturbated situations (Milligan, 1980). All these problems have led to considerable confusion about the efficacy of cluster analysis in general.

Requests for reprints should be sent to Wolfgang Schneider, who is now at the Max-Planck-Institute for Psychological Research, Leopoldstr. 24, D-8000 München 40, West Germany

Although validation research is essential in order to determine which clustering algorithms are best suited to specific applications, very little research has been devoted to this task. During the 1970s, a few Monte Carlo studies generated artificial data sets with known cluster structure as an aid in validating clustering algorithms. In particular, the mixture model (Blashfield, 1976; Wolfe, 1970) has been proposed as a useful approach for evaluating clustering procedures. According to this model, any given data set is composed of subsets of data from different populations. The task of cluster analysis is to reconstruct the true cluster structure, that is, to resolve the mixture of populations into its components. While this cannot be done with empirically derived data sets for which the number of populations and the distributional parameters of the populations are unknown, the use of Monte Carlo techniques permits the comparison of cluster analysis solutions with the known classificatory structure of the data sets. The degree of agreement between the obtained clusters and the underlying populations indicates the accuracy of the clustering solution.

Most of the mixture model studies have investigated agglomerative hierarchical methods which produce as many solution partitions as there are entities in the data set. Only a few studies (Bayne, Beauchamps, Begovich, & Kane, 1980; Blashfield, 1977; Mezzich, 1978; Milligan, 1980) have additionally evaluated nonhierarchical methods which produce only a single partition solution (see Anderberg, 1973; Cormack, 1971; Everitt, 1974; or Hartigan, 1975, for a detailed description of these clustering procedures). Since a fairly extensive review of previous Monte Carlo studies has been given elsewhere (Milligan, 1981), only the main findings will be summarized here. According to Milligan (1981), it makes sense to differentiate three different epochs of Monte Carlo studies. In the early period ranging from 1971 to 1975, researchers mainly concentrated on hierarchical methods. There was a tendency towards using only a few data sets or—in case of a more extensive sampling of the data sets—only one or two algorithms. No single procedure turned out to dominate the others. On the other hand, the results of several large-scale experiments published during the second epoch, between 1975 and 1978 (Blashfield, 1976; Kuiper & Fisher, 1975; Mojena, 1977), led to the conclusion that Ward's method using Euclidean distances seemed to outperform the other hierarchical procedures with regard to several types of data structures.

This impression, however, had to be revised when the results of the third epoch, that is, of recent Monte Carlo studies published between 1978 and 1980, were taken into account. These studies

particularly differed from those of the second period in that the problem of outliers (mini-clusters) and the choice of similarity measures were more carefully analyzed (e.g., Blashfield & Morey, 1980; Edelbrock, 1979; Edelbrock & McLaughlin, 1980), and in that a comparison of hierarchical and nonhierarchical methods was attempted (Bayne et al., 1980; Milligan, 1980). Thus Edelbrock (1979) could demonstrate that the requirement of 100% coverage of classifications in the previous Monte Carlo studies (which is actually not relevant in many applications) leads to an underestimation in those clustering procedures which are especially vulnerable to the effects of outliers. In his study, most of the hierarchical procedures performed fairly well when only 80% or 90% of the sample was to be classified. Furthermore, the studies of Edelbrock (1979) and Edelbrock and McLaughlin (1980) showed that the average linkage method and the centroid method were at least as accurate as Ward's method when correlations were used instead of Euclidean distances. In addition, comparisons of hierarchical and nonhierarchical procedures suggested that certain nonhierarchical methods were superior to any hierarchical procedure used in these studies (Bayne et al., 1980; Milligan, 1980). Although many inconsistencies in the findings make it impossible to determine which specific algorithm will be most accurate under a specified set of conditions, the results of the more recent studies indicate that a few hierarchical procedures (i.e., Ward's method; group average linkage) and some nonhierarchical methods (k-means algorithms) can be considered fairly robust and valid statistical tools.

Limitations of Previous Monte Carlo Studies

As has already been pointed out elsewhere (Edelbrock & McLaughlin, 1980; Milligan, 1981), the different methodologies and evaluative criteria used in the validation studies have made it difficult to compare findings from different mixture model experiments directly. The choice of methods for calculating accuracy of the cluster solutions (Cohen's kappa vs. Rand's statistic) appear to be of minor importance, since these measures have been shown to be highly correlated (Edelbrock & McLaughlin, 1980).[1] On the other hand the use of a variety of unrelated strategies in constructing the covariance

[1] It should be noted, however, that more recent work in this area (e.g., Fowlkes and Mallows, 1983; Milligan & Schilling, 1985) demonstrated differences among accuracy criteria and particular problems with the Rand index.

structures in the mixture model studies appears to be primarily responsible for the inconsistent outcomes. Due to different concepts of cluster structure, several researchers have either used mixtures of multivariate normal populations (e.g., Blashfield, 1976; Edelbrock, 1979; Edelbrock & McLaughlin, 1980; Kuiper & Fisher, 1975) or the concept of an ultrametric space (Cunningham & Ogilvie, 1972; Milligan & Isaac, 1980) as the basis for the generation of cluster structures. In addition, a variety of visual, spatial, and other conceptualizations was being used (Everitt, 1974; Mezzich, 1978; Mojena, 1977). Although Milligan (1981) carefully analyzed several possible sources of inconsistency in the results of Monte Carlo studies, the importance of the type of mixture for the effectiveness of different similarity coefficients has apparently been underestimated. For example, one of the most impressive findings of the more recent studies concerns the superiority of hierarchical algorithms when correlation coefficients were used instead of Euclidean distances (Edelbrock, 1979; Edelbrock & McLaughlin, 1980). A closer look at the results, however, reveals that this finding was only obtained when unconstrained mixtures of multivariate normal populations were analyzed. In particular, both Edelbrock (1979) and Edelbrock and McLaughlin (1980) reported better results for correlation coefficients when subsets of the multivariate normal mixtures developed by Blashfield (1976) were used. In the latter study, however, Euclidean distances produced more accurate solutions when multivariate gamma mixtures developed by Mojena (1977) were used. The same was true for a small Monte Carlo experiment by Milligan (1981); recovery was uniformly superior for the Euclidean distance measure. It should be noted that constrained multivariate normal mixtures were used in the Milligan experiment. That is, the distribution of the data for each cluster was truncated to prevent overlapping cluster structures, an undesirable result which certainly cannot be avoided in the unconstrained mixture approach (Milligan & Isaac, 1980). In sum, all these findings indicate that the postulated positive effects of the use of correlation coefficients seem to be confined to unconstrained mixture models.

In the present investigation, a different version of constrained mixtures of multivariate normal populations was used to test the assumption that positive effects of the use of correlation coefficients are restricted to specific mixture characteristics. In particular, the accuracy of a variety of hierarchical and nonhierarchical methods was tested using either correlation coefficients or Euclidean distances. The comparison of hierarchical and nonhierarchical methods was considered to be a major part of the experiment because the lack of empirical

evidence is especially evident here. As noted earlier, only three studies (Bayne et al., 1980; Mezzich, 1978; Milligan, 1980) have tried to relate the results of both types of clustering algorithms. Unfortunately, both Bayne et al. (1980) and Mezzich (1978) generated only two artificial data sets when comparing several hierarchical and nonhierarchical algorithms. Furthermore, Mezzich did not include the two most accurate hierarchical techniques, namely, the average linkage (or group average) method and Ward's method. The only representative comparison was done by Milligan (1980), who tested eleven hierarchical and four nonhierarchical methods by using 108 truncated multivariate normal mixtures. Milligan's mixture approach prevented the generation of overlapping cluster structures and thus provided the clustering algorithms with principally solvable tasks. Apparently, however, the precautions taken to prevent cluster overlap resulted in a particularly easy task. With regard to error-free data sets, nearly perfect recovery was obtained for all clustering algorithms. Although different types of error perturbation led to decreases in average recovery, the accuracy values remained fairly high for most procedures. In view of the high means and the small range of obtained recovery values, one is led to conclude that the results of this comparison are valid only for easily decomposable mixtures.

In the present study, an attempt was made to provide a more difficult test for the chosen clustering algorithms by using a constrained mixture approach that excluded both extremely easy and extremely difficult data sets (i.e., overlapping cluster structures) from the analysis. The generation of truncated multivariate normal mixtures very similar to those developed by Blashfield (1976) was considered important for two reasons. First, several recent Monte Carlo studies (e.g., Blashfield, 1977; Edelbrock, 1979; Edelbrock & McLaughlin, 1980) have been based on subsets of the Blashfield (1976) mixtures. This limitation to "benchmark" data sets makes it difficult to generalize the findings. The mixtures used in the present study differed from those generated by Blashfield (1976) only in that overlapping cluster structures were excluded and error perturbations were omitted in the design. In our opinion, these data sets can be regarded as a "near generalization task" for the clustering algorithms. That is, they provide an opportunity for evaluating the effects of slight changes in the data generation process.

Second, the use of nonoverlapping cluster structures in the present study makes it also possible to test Milligan's (1981) interesting hypothesis concerning the effect of cluster overlap on the efficiency of hierarchical methods. According to this assumption, Ward's method

gives the best recovery only when overlap is present in the cluster structure. On the other hand, the group average or average linkage method is supposed to give superior recovery with non-overlapping structures. Thus, the group average method was expected to be the most efficient hierarchical algorithm in the present study.

In sum, the major purpose of this study was to compare the accuracies of a variety of hierarchical and nonhierarchical clustering techniques at different coverage levels. In addition, the impact of different measures of similarity (Euclidean distances vs. Pearson correlation coefficients) on the accuracy of the clustering algorithms was assessed. To overcome some of the problems (i.e., cluster overlap) of previous Monte Carlo studies using unconstrained multivariate normal mixtures, truncated multivariate normal mixtures were generated in the present study which satisfied the requirements of external isolation and internal cohesion of the resulting clusters (see Cormack, 1971). That is, entities in one cluster were similar to each other (internal cohesion) and did not overlap with entities in other clusters (external isolation).

Method

Data Sets

Each data set was a mixture of samples drawn from a number of different populations. As has already been noted by Blashfield (1976), two kinds of parameters can be distinguished from each data set, namely, (1) parameters of the populations represented in the mixture, and (2) the parameters of the mixture itself.

Three different types of population parameters were considered, namely, the means and the variances of the population of the variables, and the correlational structure among the variables. The means of the populations of the variables were chosen randomly from a uniform distribution ranging from 45.0 to 60.0, while the variance for each variable was chosen from a uniform distribution which ranged from 5.0 to 30.0. The correlation structure among the variables was specified in the same way as in the Blashfield (1976) study. That is, the number of principal components ranged from 2 to 10 and was always less than the number of variables. The variables were assumed to have a multinormal distribution. The number of entities sampled from each population was determined by randomly choosing an integer from a uniform distribution which ranged from 5 to 50.

With regard to the parameters of the mixture, the number of populations represented ranged from 2 to 6, and the number of variables ranged from 3 to 25. Both of these parameters were determined by randomly choosing an integer from a uniform distribution with the specified ranges. The total number of entities in the mixtures ranged from 18 to 265.

The data sets differed from those generated by Blashfield (1976) in that no measurement error was added to the data points. According to Blashfield, measurement error was included in his study to increase the difficulty of the mixtures, that is, to insure that obtaining good recovery values would not be a trivial result. In our view, the addition of measurement error creates a problem, in that the task for the clustering algorithms is to guess the "true" values. In the present study, the mixtures were made adequately difficult by excluding extremely easy data sets (to be defined below) from the analysis instead.

Clustering Algorithms

The simulated data sets were clustered by the nine agglomerative hierarchical algorithms and the four nonhierarchical procedures listed in Table 1. The hierarchical clustering techniques have been described and discussed in detail elsewhere (Anderberg, 1973; Cormack, 1971). In brief, these methods are iterative and generate solutions which can be graphically represented as hierarchical trees (dendrograms). Each level of the tree represents a different clustering called a partition. If a sample consists of N entities, then the first partition will yield $N - 1$ clusters, the next $N - 2$ clusters, and so on up to the last partition which consists of one cluster containing all individuals. According to Johnson (1967), the only difference between agglomerative hierarchical clustering algorithms concerns the step in which the distance between the new cluster and the remaining data points is computed. According to Lance and Williams (1967) and Wishart (1969), a general equation ("recurrence formula") describes how the various hierarchical algorithms compute this distance:

$$[1] \qquad d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \lambda \, |d_{hi} - d_{hj}|.$$

In this equation, d_{ij} denotes the Euclidean distance between the entities i and j which have been combined to form a new cluster k. The Euclidean distance between remaining entities h and the new cluster k is denoted by d_{hk}. The values of the parameters α_i, α_j, β, and λ depend on the particular clustering method and are given in Table 1.

Table 1

Cluster Analysis used in the present Monte Carlo Study and their Parameter Values in the Recurrence Formula

Method	Parameter Values			
	α_i	α_j	β	λ
a) Hierarchical				
(1) Single Linkage	0.5	0.5	0	-0.5
(2) Complete Linkage	0.5	0.5	0	0.5
(3) Average Linkage	n_i/n_k	n_j/n_k	0	0
(4) Median	0.5	0.5	0.25	0
(5) Centroid	n_i/n_k	n_j/n_k	$-n_in_j/n^2_k$	0
(6) Ward's Method)	$(n_h+n_i)/(n_h+n_k)$	$(n_h+n_j)/(n_h+n_k)$	$-n_h/n_h+n_k$	0
(7) Beta-Flexible I	0.75	0.75	0.5	0
(8) Beta Flexible II	0.625	0.625	0.25	0
(9) McQuitty's Method	0.5	0.5	0	0
b) Nonhierarchical				
(1) CLUSTAN k-means: Every k^{th} element				
(2) CLUSTAN k-means: Ward's Centroid				
(3) Spaeth's k-means: Every k^{th} element				
(4) Spaeth's k-means: Ward's Centroid				

n_i = number of entities in cluster i of preceding partition
n_j = number of entities in cluster j of preceding partition
n_k = number of entities in cluster i or cluster j of preceding partition
n_h = number of remaining entities for which the distance to cluster k has to be recomputed

Two different versions of the flexible-beta method (Lance & Williams, 1967) were included to assess the impact of changes in beta on average recovery values. That is, in addition to the beta-value recommended by Lance and Williams ($\beta = -.25$), further analyses were run for $\beta = -.5$. (It should be mentioned that McQuitty's method can also be regarded as a version of the flexible-beta method where β equals zero.) As an earlier preliminary study (Scheibler & Schneider, 1978) has shown, the recommended β-value may not necessarily lead to the best recovery values for this procedure.

The nonhierarchical procedures used in the present study have the common feature that they produce only a single partition solution. That is, the researcher usually specifies in advance which number of clusters should build up the solution (see Anderberg, 1973, for a more detailed description). As Blashfield (1977) and Milligan (1980) have emphasized, nonhierarchical procedures seem to be sensitive to the nature of the starting partition. For the purpose of the present study, two different starting procedures were selected for two nonhierarchical (k-means) methods, namely, the CLUSTAN k-means algorithm (Wishart, 1975) and the k-means algorithm offered by Späth (1975, 1980). Both randomly selected data units and the centroids obtained from Ward's method were used as starting cluster centroids for these two iterative partitioning algorithms. Although the CLUSTAN k-means algorithm and Späth's k-means approach are comparable in that they rely on MacQueen's method (Anderberg, 1973), they differ in the following respect: While in the CLUSTAN procedure each cluster is represented by its centroid, in Späth's algorithm each cluster is represented by its median.

Procedure

All computations were based on a total of 200 constrained random mixtures. In generating these data sets, all those mixtures for which discriminant analyses yielded imperfect solutions were eliminated. As noted by Milligan and Isaac (1980), a nonzero misclassification rate with a linear discriminant function analysis indicates an overlapping cluster structure. Similarly, all data sets which could be easily reconstructed by at least one of the algorithms under investigation were excluded from further analysis. Because results of a previous Monte Carlo study (Scheibler & Schneider, 1978) using 766 mixtures had shown that Ward's method did always belong to the group of algorithms that came to perfect solutions whenever those solutions were obtained, Ward's method was chosen to minimize bias against any method. The two-step procedure included in the data generation process allowed us to eliminate extremely easy and extremely difficult data sets until a total of 200 constrained random mixtures was available for analysis.

For each mixture, different parameter values were specified. The 13 clustering algorithms were applied to standardized versions of all mixtures. The standardization for Späth's procedure differed from the others (i.e., z-transformations) in that linear transformations were done so that equal ranges and equal minimum values and maximum

values resulted for all variables. According to Späth (1975, 1980), this standardization method is superior to many others.

Euclidean distances and Pearson correlation coefficients were used as similarity measures for all algorithms. The same computer program, CLUSTAN 1C (Wishart, 1975) was used for all hierarchical methods and also for the first two nonhierarchical procedures. Späth's k-means analyses were performed using a FORTRAN IV program written by Späth.

The statistic kappa (Cohen, 1960) was used to assess the accuracies of the cluster solutions. Kappa values range from -1 to 1, with larger values indicating larger levels of agreement between the obtained clusters and the underlying populations. In order to assess the effects of different levels of coverage on the recovery values of the clustering algorithms, a procedure slightly different from that of Edelbrock and McLaughlin (1980) was chosen. Edelbrock and McLaughlin first calculated accuracies at 100% coverage and then proceeded to successively lower levels in the hierarchical tree (i.e., 99, 95, 90, 80, 70, 60, 50, and 40 percent coverage). In the present study, 10 different solutions were obtained for each clustering algorithm. First, a solution containing $k + 9$ clusters (k = number of populations represented in the mixture) was analyzed. The 9 smallest clusters were excluded from analysis, and only the k largest clusters were compared with the k populations represented in the mixture. The number of mini-clusters was successively reduced in the following 9 steps. In the last step, the number of clusters equalled the number of populations (i.e., 100% coverage). Thus, successively more mini-clusters were included in the computation of the recovery values. At each step, kappa values were obtained for all algorithms, and the whole procedure was repeated 200 times. Clusters were matched by analyzing all possible relationships among the given and reconstructed cluster structures via systematic permutations. Those relationships yielding the highest kappa values were chosen as optimal. The procedural details of the Monte Carlo simulations are summarized in Figure 1.

Results

An overview of the results is given in Tables 2 and 3, where the recovery values of the different clustering algorithms are compared for Euclidean distances and Pearson correlation coefficients, respectively.

In Table 2 it is easy to see that most clustering algorithms using

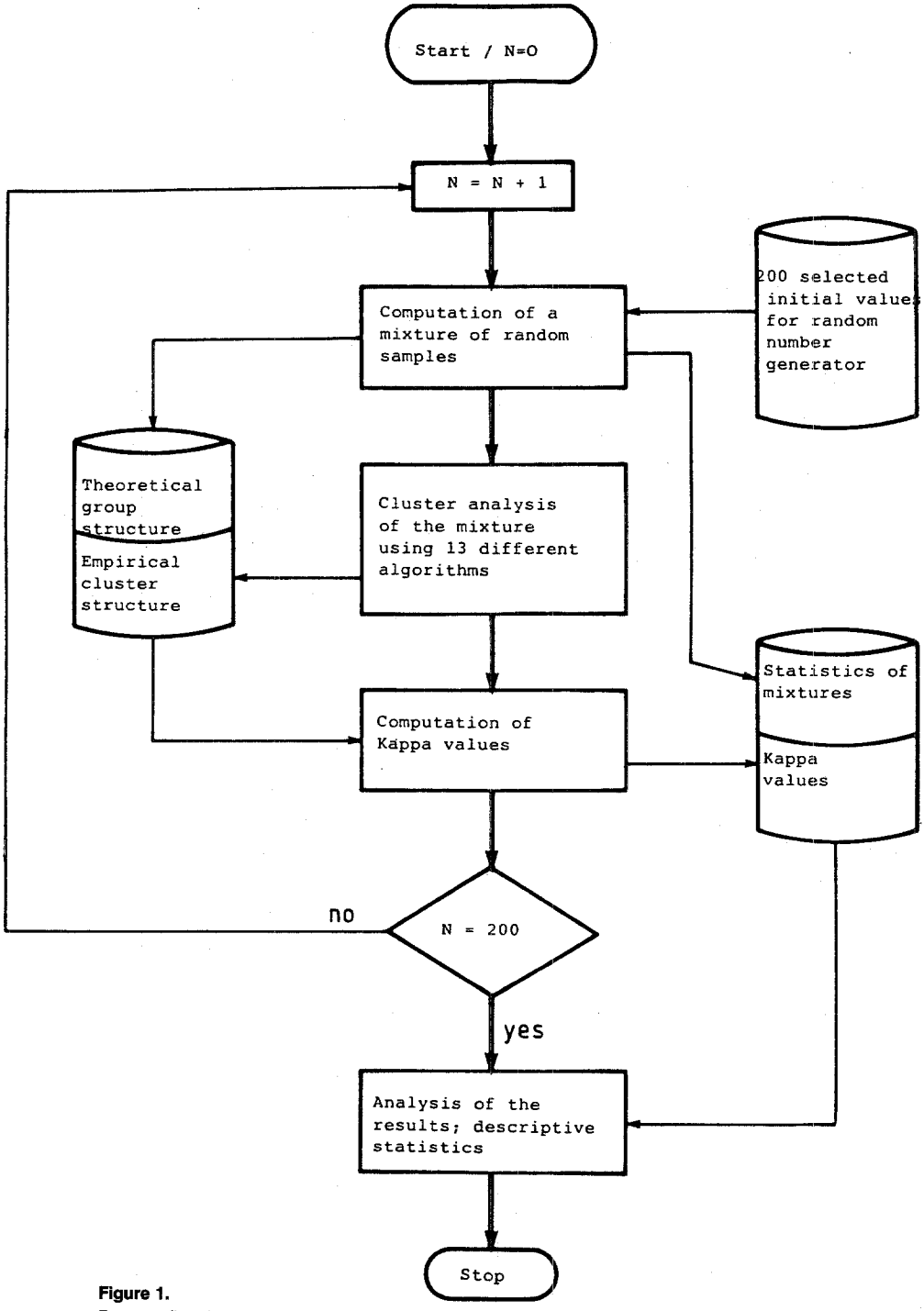

Figure 1.
Program flowchart describing the Monte Carlo test of cluster analysis algorithms.

Table 2

Median kappas (uncorrected) for the different clustering algorithms using Euclidean distances, K + 9 to K clusters (K = number of underlying populations)

	N of Clusters									
	K+9	K+8	K+7	K+6	K+5	K+4	K+3	K+2	K+1	K
Single Linkage	1.	.89	.86	.82	.79	.75	.70	.59	.50	.04
Complete Linkage	1.	1.	1.	.98	.94	.89	.82	.69	.57	.38
Average Linkage	1.	1.	1.	1.	.98	.97	.95	.90	.77	.16
Median Method	1.	1.	1.	1.	1.	.87	.82	.75	.60	.04
Centroid Method	1.	1.	1.	1.	1.	.90	.86	.80	.65	.05
Ward's Method	1.	1.	1.	1.	1.	1.	1.	.96	.90	.791
Lance-Williams' Method (beta = -.5)	1.	1.	1.	1.	1.	1.	.98	.93	.87	.77
Lance-Williams Method (beta = -.25)	1.	1.	1.	1.	1.	1.	.98	.94	.86	.72
McQuitty's Method (beta = 0)	1.	1.	1.	1.	1.	.97	.95	.90	.75	.43
Clustan's k-means (random)	1.	1.	1.	1.	1.	.98	.94	.88	.81	.67 (.775)
Späth's k-means (random)	.98	.95	.95	.92	.91	.86	.85	.77	.69	.55 (.773)

Note. Numbers in parentheses refer to solutions for nonrandom starting seeds.

Table 3

Median kappas (uncorrected) for the different clustering algorithms using Pearson correlation coefficients for K + 9 to K cluster (K = number of underlying populations)

	N of clusters									
Method	K+9	K+8	K+7	K+6	K+5	K+4	K+3	K+2	K+1	K
Single Linkage	1.	1.	1.	1.	.98	.98	.97	.92	.80	.43
Complete Linkage	1.	1.	.98	.95	.90	.87	.80	.72	.63	.49
Average Linkage	1.	1.	1.	1.	1.	.99	.98	.96	.92	.81
Median Method	1.	1.	1.	1.	1.	.99	.97	.95	.89	.75
Centroid Method	1.	1.	1.	1.	.99	.98	.95	.90	.83	.70
Ward's Method	1.	1.	1.	1.	.98	.96	.95	.91	.88	.78
Lance-Williams' Method (beta = -.5)	1.	1.	1.	1.	1.	.98	.96	.93	.87	.73
Lance-Williams Method (beta = -.25)	1.	1.	1.	1.	1.	.98	.96	.94	.88	.75
McQuitty's Method (beta = 0)	1.	1.	1.	1.	1.	1.	.97	.94	.87	.73
Clustan's k-means (random)	1.	1.	1.	.99	.98	.96	.92	.88	.80	.66 (.72)

Note. The number of parentheses refers to the solution for nonrandum starting seeds.

Euclidean distance measures performed well when there were more clusters than underlying populations, that is, when only a certain percentage of the sample had to be classified. As already noted by Edelbrock (1979), the largest decrease in accuracy was generally observed from level $k + 1$ to level k, where all elements had to be assessed. The drop in accuracy is probably due to the effects of chaining, since those algorithms known to be highly susceptible to chaining (i.e., single linkage, median method, centroid method) show the most impressive decrease in accuracy when outliers have to be included in the analysis.

Totally different results were obtained when the Pearson correlation coefficient was used as a measure of similarity (cf. Table 3). Summing across clustering algorithms, recovery values were significantly better ($p \leqq .0001$) for correlation coefficients, compared with Euclidean distances (cf. the similar findings by Edelbrock, 1979).

It can be seen that the decrease in accuracy is approximately the same for both similarity measures when only the first 5 or 6 levels of coverage are analyzed. Undoubtedly, the most remarkable difference in results is obtained for level k, where all elements have to be classified. While the drop in accuracy level from $k + 1$ to level k is still significant ($p \leqq .01$) when correlation coefficients are used instead of Euclidean distances, recovery values are now acceptable for most algorithms. In particular, an enormous improvement can be observed for the procedures probably susceptible to chaining (i.e., single linkage method, median method, centroid method) and for the average linkage algorithm value. Similar to the results obtained by Edelbrock (1979), the average linkage algorithm tended to be more accurate than all other algorithms when correlation coefficients were used. Although the recovery values for Ward's method were slightly lower, this method ranked second best in the correlation coefficient condition. It should be emphasized that most algorithms performed nearly almost equally well with regard to the first 9 levels of coverage (i.e., from $k + 9$ to $k + 1$), and that all algorithms except for the single linkage and complete linkage method led to acceptable recovery values when all elements had to be classified (i.e., level k).

Interestingly enough, the solutions for the two versions of the Lance-Williams method used in this study did not differ significantly for both Euclidean distances and Pearson correlation coefficients. Unexpectedly, the beta-value recommended by Lance and Williams (beta = $-.25$) did not lead to better results than the arbitrarily chosen beta-value (beta = $-.5$). This finding confirms the conclusion of Scheibler and Schneider (1978) that the range of possible beta values

should be more carefully explored with regard to its role for algorithm accuracy.

In order to permit parametric analyses of kappa, the variances of this proportion measure had to be stabilized and normalized using the arcsin transformation (Winer, 1971, pp. 399–400),

$$K^* = 2 \text{ ARCSIN } \sqrt{K},$$

where K^* is the transformed kappa value used in the analyses, and K indicates the empirically derived kappa value.

First, the transformed kappa values were analyzed using a 10 × 2 × 10 (algorithms × similarity measure × level of coverage) ANOVA with repeated measures. As Pearson correlation coefficients could not be used for Späth's k-means method, separate comparisons focussing on the Euclidean distance measure were made for Späth's k-means method and the other hierarchical and nonhierarchical procedures.

As can be seen from Table 4, several significant effects were found in the 10 × 2 × 10 ANOVA, most of them confirming the results obtained by Edelbrock (1979; Edelbrock and McLaughlin, 1980).

Table 4

Results of the 10 x 2 x 10 (Algorithms x similarity measure x level of coverage) ANOVA with repeated measures

Dependent Variables	Degrees of Freedom	F-value
Measure (M)	1;199	94.1
Algorithm (A)	9;1791	185.1
Coverage (C)	9;1791	1735.0
Interactions		
M x A	9;1791	114.9
M x C	9;1791	170.0
A x C	81;16119	59.7
M x A x C	81;16119	38.97

Note. All F-values are significant at the .001 level.

Accuracy was inversely related to coverage across algorithms. Collapsing across similarity measures, significant differences were found among the 10 hierarchical algorithms, with higher accuracies obtained for Ward's method and the Lance-Williams methods. Furthermore, the effects of the measure of similarity, the measure of similarity by clustering algorithm interaction, and the measure of similarity by clustering algorithm by level of coverage interaction all remained significant.

Additional analyses including Späth's *k*-means method and Euclidean distances as measure of similarity confirmed the results obtained by Blashfield (1977) and Milligan (1980). That is, the two *k*-means algorithms (i.e., Clustan's *k*-means and Späth's *k*-means) produced recovery values worse than those of the best hierarchical methods (i.e., Ward's method, the Lance-Williams methods) when random starting seeds were used. On the other hand, when the centroids of the clusters generated by Ward's method were used as starting seeds, Clustan's *k*-means and Späth's *k*-means produced recovery values that were equivalent to those of the best hierarchical methods (.775 and .773, respectively). This finding underlines Milligan's (1980) conclusion that the "starting partition must be close to the final solution if the *k*-means algorithms are to be expected to give good recovery" (p. 339). At the same time, it sheds doubt on Späth's hypothesis that his clustering algorithm will perform best when random starting seeds are used.

Post-hoc comparisons using the Newman-Keuls procedure revealed that—at a *p*-level of .05—the algorithms can be divided into six and five accuracy groups for Euclidean distances or correlation coefficients, respectively.

When Euclidean distances were used as similarity measure, the most accurate algorithms included the Lance-Williams method (beta = −.5), Ward's method and the *k*-means procedures using nonrandom starting seeds. The next subset was comprised of the Lance-Williams method (beta = −.5) and the *k*-means procedures using random starting seeds. The third group consisted of McQuitty's method and the complete linkage method. The three groups with lowest accuracy included the average linkage method (subset 4), centroid method (subset 5), and the single-linkage and median method (subset 6).

Post-hoc comparisons for correlation coefficients yielded a different pattern of results. There was only one algorithm—the average-linkage method—located in the highest accuracy subset. The next group included McQuitty's method, the Lance-Williams methods (beta = −.5 and beta = −.25), Ward's method, and the median method and

Clustan's k-means using random starting seeds. Finally, the two low-accuracy subsets 4 and 5 consisted of the complete-linkage and single-linkage method, respectively.

From these results, it appears that Ward's method, the Lance-Williams method (beta = $-.5$), and Clustan's k-means using nonrandom starting seeds yielded accurate solutions regardless of type of similarity and can be recommended for application. In addition, the average linkage method seems particularly appropriate when correlation coefficients are used as similarity measure. On the other hand, the single-linkage method performed poorly regardless of type of similarity measure, followed by the centroid method that also yielded low accuracy values in both conditions. Thus these two methods cannot be recommended for application.

Discussion

One of the most interesting aspects of the present study concerned the question of whether the main results of previous Monte Carlo studies evaluating clustering algorithms could be confirmed when a different data generation process was used. Taken together, the results of our simulation study appear to be encouraging. In accordance with previous investigations using either multivariate normal mixtures (Blashfield, 1977) bivariate normal mixtures (Bayne et al., 1980), or the concept of an ultrametric space (Milligan, 1980), the non-hierarchical (k-means) algorithms used in the present study were found to produce excellent recovery of cluster structure provided that the starting seeds were obtained from a robust hierarchical clustering algorithm (i.e., Ward's method). It should not be overlooked, however, that both nonhierarchical algorithms recovered true structure worse, on average, than the initial partition given to them (i.e., the solution of Ward's method). For Clustan's k-means, this was true regardless of the type of similarity/dissimilarity measure used, although the absolute differences between Ward's and k-means' solutions were small. On the other hand, the k-means algorithms yielded recovery values significantly worse than those of the two most robust hierarchical methods (i.e., Ward's method and the Lance-Williams method) when random starting seeds were used. Thus, in spite of recommendations given by some authors (e.g., Späth, 1980), evidence from different Monte Carlo simulations leads to the conclusion that the use of random starting points will generally result in suboptimal classifications.

With regard to the effect of level of coverage, the results of the

present study validate the findings by Edelbrock (1979) and Edelbrock and McLaughlin (1980). Regardless of the type of Monte Carlo data sets (i.e., constrained or unconstrained multivariate normal mixtures or multivariate gamma mixtures), the general finding was that accuracy values increased as coverage decreased, and that the largest decrease in recovery occurred when all elements had to be classified. The latter is particularly true when Euclidean distances are used as similarity measure, thus confirming the results by Edelbrock (1979).

Contrary to expectations, most hierarchical algorithms performed significantly better when correlation was used as a measure of similarity. According to our hypothesis, the positive effects of the use of correlation coefficients should be confined to unconstrained multivariate normal mixtures of the type used by Blashfield (1976), Edelbrock (1979; Edelbrock & McLaughlin, 1980), and Milligan (1981). However, the use of constrained multivariate normal mixtures in our study did not yield a different pattern of results. Obviously, simply excluding cluster overlap does not necessarily change the effects of the similarity measure. The superiority of correlation coefficients to Euclidean distances may be due to the different treatment of outliers in the data sets. As Edelbrock (1979) has pointed out, correlation has a limited range of similarity/dissimilarity not sensitive to elevation. Consequently, here outliers are probably more similar to other elements than in analyses using Euclidean distances, and thus are combined into clusters at lower levels in the hierarchy. The hypothetical example in Figure 2 may be helpful in understanding why most clustering algorithms were more accurate when correlations were used as similarity measure. In Figure 2, d_{12} denotes the Euclidean distance between data points x_1 and x_2, whereas the correlation between these data points is defined by the cosine of the angle between them (i.e., $r_{12} = \cos \theta_{12}$). Similarly, d_{34} denotes the Euclidean distance between data points x_3 and x_4, and $\cos \theta_{34}$ defines the correlation between these data points.

Consider that the single linkage method and Euclidean distances were used in the example, and that the elements of $K = 3$ populations had to be reconstructed. Due to the chaining tendency of the single linkage procedure (see Anderberg, 1973), there is a high probability that samples A and B would be agglomerated at an early hierarchical level. Further, data point x_1 would be treated as an outlier and build up a cluster of its own. As a consequence, only sample C would be correctly reconstructed and low kappa values would result.

On the other hand, when the single linkage method and correlation coefficients were used in the example given in Figure 2, a correct

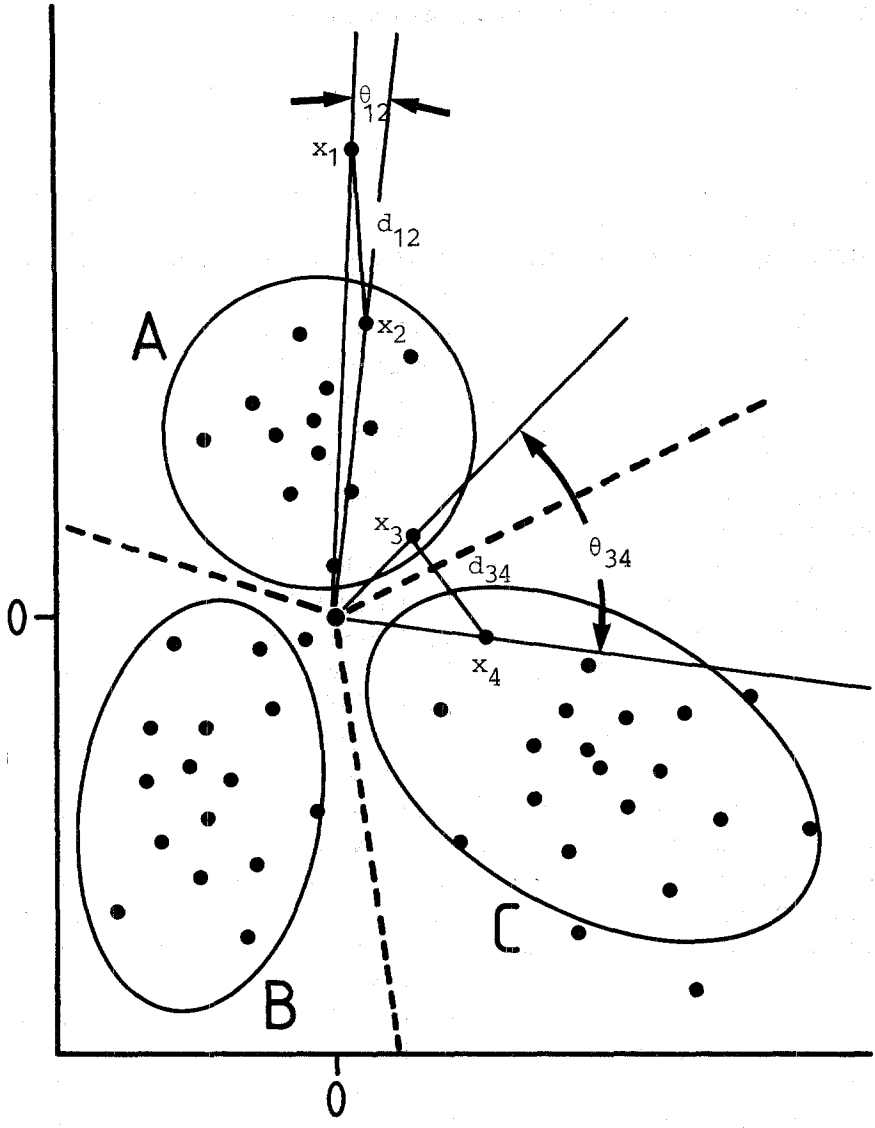

Figure 2.
Hypothetical example of a sample mixture out of three populations in a two-dimensional (orthogonal) data space illustrating different results for clustering algorithms using either Euclidean distances or correlation coefficients.

solution was obtained. This is due to the fact that the angles between the elements within samples A, B, and C are relatively small (and the corresponding correlations relatively high), but large between elements from different samples (e.g., angle θ_{34}).

The choice of non-overlapping mixtures is probably of importance

here, because, for overlapping mixtures, it is possible to obtain large angles (up to 180°) between two elements within a sample. But given the superior performance of correlation coefficients in earlier studies using unconstrained mixtures (e.g., Edelbrock, 1979), the mixture generation process cannot solely account for the effects of the type of similarity measure.

The results of the present study and that of Edelbrock (1979) suggest that correlation may be more appropriate as a measure of similarity than Euclidean distances. This does not mean, however, that the use of correlations should be recommended for all applications; rational considerations are always important, and the choice of the similarity measure should best be dictated by the nature of the data. Thus, for example, since it may be true that our simulation created data where level is relatively unimportant relative to shape and scatter, the simulation results may not be generalizable to data where level is a central feature.

Finally, the results of the present study may be used to evaluate the validity of Milligan's (1981) conclusions regarding the factors determining accuracy of Ward's method and the average linkage algorithm (i.e., the two best hierarchical methods). Based on his review of Monte Carlo tests of cluster analysis, Milligan inferred that three alternative factors appear to determine when Ward's method or the group average technique provide best recovery:

(1) the selection of similarity measure; Ward's method is expected to give better recovery for Euclidean distances while group average procedures provide equivalent accuracy when Pearson correlation coefficients are used;
(2) the treatment of outliers—Ward's method is supposed to be superior when total coverage is required, while the group average procedure gives at least equivalent results when not all elements have to be classified;
(3) the influence of cluster overlap—Ward's method works better when cluster overlap is provided, while the group average method gives superior recovery when data do not possess overlapping structure.

Undoubtedly, conclusion (3) was not confirmed in the present study, that is, Ward's method ranked first in the Euclidean distances condition although nonoverlapping data structures were used. On the other hand, conclusion (1) could be completely verified, as the interaction between similarity measures and algorithm performance proved to be significant. With regard to conclusion (2), a more differentiated formulation seems to be necessary; the conclusion was confirmed when

Euclidean distances were used as the measure of similarity but was not correct for correlation coefficients. Here, the group average method proved at least equivalent regardless of level of coverage.

In sum, the comparison of a broad range of hierarchical and nonhierarchical clustering algorithms using truncated multivariate normal mixtures led to two main conclusions. First, most results obtained for hierarchical algorithms with unconstrained multivariate normal mixtures could be generalized to truncated multivariate normal data sets. That is, most algorithms found to be fairly accurate in previous studies (e.g., Ward's method, group average procedure, the Lance-Williams method) also ranked high in the present evaluation analysis, and vice versa. In addition, the effects of measure of similarity and level of coverage were comparable to those found in previous studies.

Second, the comparison of hierarchical and nonhierarchical clustering algorithms showed that the latter were equally effective as the most robust hierarchical ones provided that nonrandom starting seeds were used. Thus, by and large, a small group of hierarchical as well as nonhierarchical cluster algorithms could be identified that proved to be fairly accurate. Nonetheless, caution must be exercised when attempting to generalize these findings. Future mixture model studies should include mixtures whose multivariate distributions are a long way from normality in order to systematically explore possible effects due to type of mixture. Although no single algorithm appears best for all applications, a more extended analysis of mixture characteristics should help to identify a core group of algorithms preferable for a broad variety of clustering problems.

References

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

Bartko, J. J., Strauss, J. S., & Carpenter, W. T. (1971). An evaluation of taxometric techniques for psychiatric data. *Classification Society Bulletin, 2,* 2–8.

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.

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

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

Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature on 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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20,* 37–46.
Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society (Series A), 134,* 321–367.
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.
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.
Everitt, B. S. (1974). *Cluster analysis.* London: Heinemann.
Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder). *Journal of the American Statistical Association, 78,* 553–584.
Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.
Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32,* 241–254.
Kuiper, F. K., & Fisher, L. A. (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 system. *Computer Journal, 9,* 373–380.
Mezzich, J. (1978). Evaluating clustering methods for psychiatric diagnosis. *Biological Psychiatry, 13,* 265–281.
Mezzich, J., & Solomon, H. (1980). *Taxonomy and behavioral science—Comparative performance of grouping methods.* New York: Academic Press.
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. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16,* 379–407.
Milligan, G. W., & Isaac, P. D. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition, 12,* 41–50.
Milligan, G. W., & Schilling, D. A. (1985). Asymptotic and finite sample characteristics of four external criterion measures. *Multivariate Behavioral Research, 20,* 97–109.
Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20,* 359–363.
Scheibler, D., & Schneider, W. (1978). *Probleme und Ergebnisse bei der Evaluation von Clusteranalyse-Verfahren* [Evaluation of clustering algorithms]. Bericht aus dem Psychologischen Institut der Universität Heidelberg, No. 11.
Späth, H. (1975). *Cluster-Analyse-Algorithmen zur Objektklassifikation und Datenreduktion* [Clustering algorithms for object classification and data reduction]. Muenchen: Oldenbourg.
Späth, H. (1980). *Cluster analysis algorithms.* Chichester, England: Ellis Horwood.
von Eye, A., & Wirsing, M. (1978). An attempt for a mathematical foundation and evaluation of MACS, a method for multidimensional automatical cluster detection. *Biometrical Journal, 20,* 655–666.
von Eye, A., & Wirsing, M. (1980). Cluster search by enveloping space density maxima. In M. M. Barritt & D. Wishart (Eds.), *COMPSTAT 1980.* Vienna: Physica-Verlag.
Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw Hill.
Wishart, D. (1969). An algorithm for hierarchical classifications. *Biometrics, 25,* 165–170.
Wishart, D. (1975). *CLUSTAN 1C user manual.* London: Computer center.
Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5,* 329–350.