PSYCHOMETRIKA—VOL. 50, NO. 2, 159–179
JUNE 1985

AN EXAMINATION OF PROCEDURES FOR DETERMINING THE NUMBER OF CLUSTERS IN A DATA SET

GLENN W. MILLIGAN AND MARTHA C. COOPER
THE OHIO STATE UNIVERSITY

A Monte Carlo evaluation of 30 procedures for determining the number of clusters was conducted on artificial data sets which contained either 2, 3, 4, or 5 distinct nonoverlapping clusters. To provide a variety of clustering solutions, the data sets were analyzed by four hierarchical clustering methods. External criterion measures indicated excellent recovery of the true cluster structure by the methods at the correct hierarchy level. Thus, the clustering present in the data was quite strong. The simulation results for the stopping rules revealed a wide range in their ability to determine the correct number of clusters in the data. Several procedures worked fairly well, whereas others performed rather poorly. Thus, the latter group of rules would appear to have little validity, particularly for data sets containing distinct clusters. Applied researchers are urged to select one or more of the better criteria. However, users are cautioned that the performance of some of the criteria may be data dependent.

Key words: classification, stopping rules, numerical taxonomy.

Introduction

In most real life clustering situations, an applied researcher is faced with the dilemma of selecting the number of clusters or partitions in the final solution (Everitt, 1979; Sneath & Sokal, 1973). Virtually all clustering procedures provide little if any information as to the number of clusters present in the data. Nonhierarchical procedures usually require the user to specify this parameter before any clustering is accomplished and hierarchical methods routinely produce a series of solutions ranging from n clusters to a solution with only one cluster present (assume n objects in the data set). As such, numerous procedures for determining the number of clusters in a data set have been proposed (Dubes & Jain, 1979; Milligan, 1981c; Perruchet, 1983). When applied to the results of hierarchical clustering methods, these techniques are sometimes referred to as stopping rules. Often, such rules can be extended for use with nonhierarchical procedures as well.

The application of a stopping rule in a cluster analytic situation can result in a correct decision or in a decision error. Basically, two different types of decision errors can result. The first kind of error occurs when the stopping rule indicates k clusters are present when, in fact, there were less than k clusters in the data. That is, a solution containing too many clusters was obtained. The second kind of error occurs when the stopping rule indicates fewer clusters in the data than are actually present. Hence, a solution with too few clusters was obtained. Although the severity of the two types of errors would change depending on the context of the problem, the second type of error might be considered more serious in most applied analyses because information is lost by merging distinct clusters.

The present study reports the results of a simulation experiment designed to determine the validity of 30 stopping rules already in the clustering literature. Although a vast number of references exist, few comparative studies have been performed on these mea-

The authors would like to express their appreciation to a number of individuals who provided assistance during the conduct of this research. Those who deserve recognition include Roger Blashfield, John Crawford, John Gower, James Lingoes, Wansoo Rhee, F. James Rohlf, Warren Sarle, and Tom Soon.

Requests for reprints should be sent to Glenn W. Milligan, Faculty of Management Sciences, 301 Hagerty Hall, The Ohio State University, Columbus, OH 43210.

0033-3123/85/0600-6014$00.75/0

©1985 The Psychometric Society

sures. Authors continue to introduce new stopping criteria while providing little or no comparative performance information. Since the stopping rules are heuristic, ad hoc procedures, an applied researcher must critically examine the suggested solution provided by any such index. The present simulation results should help applied researchers in this evaluation task.

The remainder of the paper is organized in the traditional method, results, and discussion sections format. In the methods section, a description of the test data sets is provided along with a discussion of the 30 stopping rules. The results section presents the findings of the simulation study. Finally, the discussion section interprets the results, offers recommendations for the application of the stopping rules in applied situations, and gives suggestions for continued research.

Method

Data Sets

The artificial data sets used in the present study contained either 2, 3, 4, or 5 distinct nonoverlapping clusters. The data sets consisted of a total of 50 points each and the clusters were embedded in either a 4, 6, or 8 dimensional Euclidean space. Overlap of cluster boundaries was not permitted on the first dimension of the variable space. The absolute minimum separation between neighboring cluster boundaries on the first dimension was equal to .25 times the sum of the within-cluster standard deviations from the two respective clusters. The actual distribution of the points within clusters followed a (mildly) truncated multivariate normal distribution. Hence, the resulting structure could be considered to consist of "natural" clusters which exhibited the properties of external isolation and internal cohesion.

Since a larger number of dimensions tended to contain more (redundant) information as to the clustering in the data, cluster recovery by the methods tended to increase with increasing dimensionality. Similar results were found for the best stopping rules. That is, the better rules capitalized on this redundant information and thus exhibited greater accuracy. The poorer rules tended to display fairly constant recovery as the number of dimensions increased.

The design factors corresponding to the number of clusters and to the number of dimensions were crossed with each other and both were crossed with a third factor that determined the number of points within the clusters. This third factor consisted of three levels where one level had an equal number of points in each cluster (or as close to equality as possible). The second level required that one cluster must always contain 10% of the data points, whereas the third level required that one cluster must contain 60% of the items. The remaining points were distributed as equally as possible across the other clusters present in the data. The 60% condition produced a marked discrepancy in cluster sizes for data sets with larger number of clusters while the 10% condition produced a discrepancy when few clusters were present. Overall, there were 36 cells in the design. Three replications were generated in each cell of the design. This produced 108 data sets for testing purposes. Each data set was used to compute a dissimilarity matrix consisting of the Euclidean distances between points. The matrix of distances was the input data for the clustering methods. Each matrix was analyzed by the four clustering methods to provide a variety of solutions. Thus, the design produced a total of 432 test solutions. The four hierarchical clustering methods used to generate the solutions were the single link, complete link, group average, and Ward's minimum variance procedures. The data generation process used in the present experiment corresponds to the error-free data conditions in previous studies (Milligan 1980; Milligan, 1981b).

It is useful to stress the fact that the clusters were internally cohesive and well sepa-

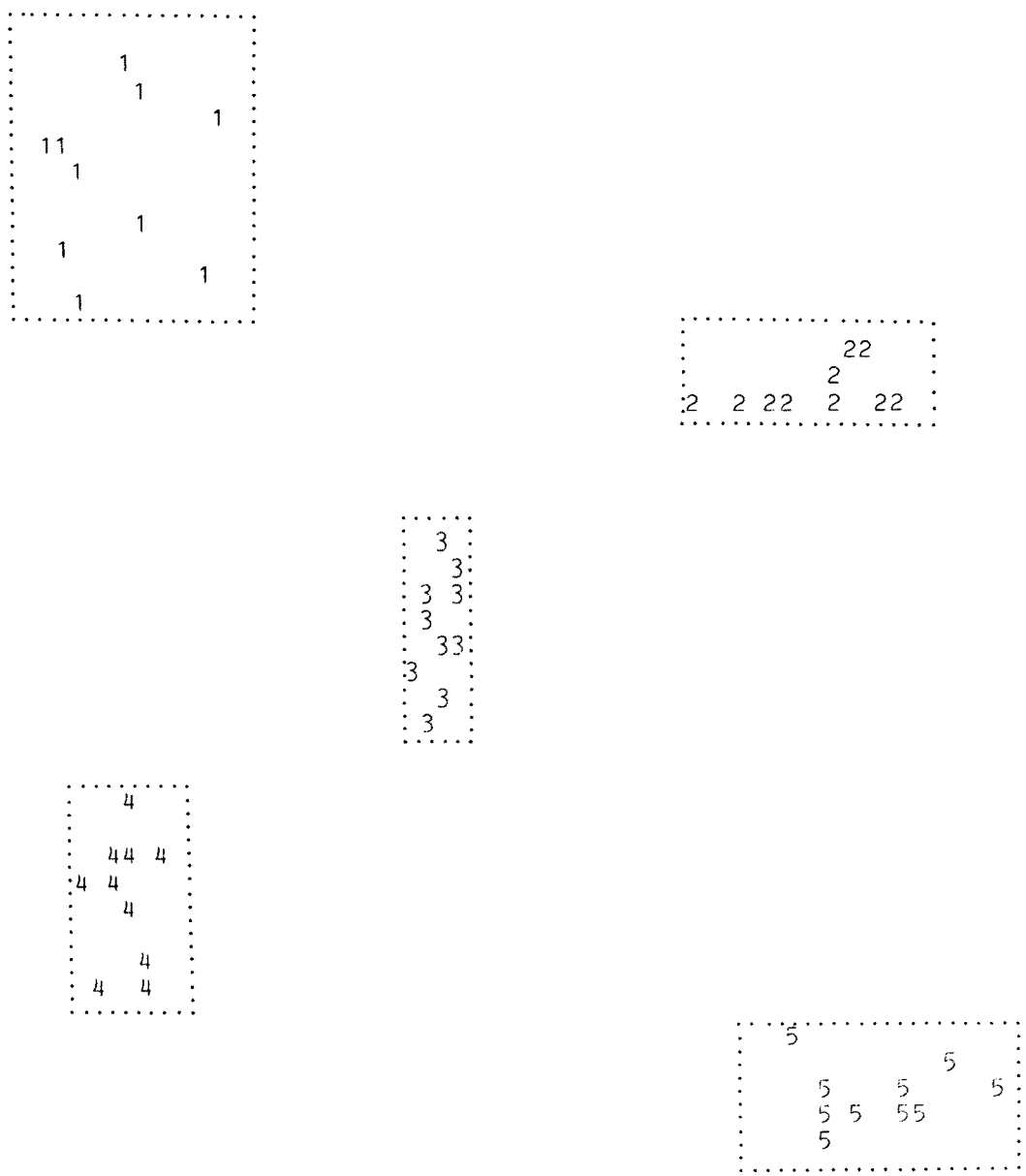


FIGURE 1
Two-dimensional representation of a five cluster data set with 10 points per cluster.

rated in the variable space. This fact can be seen in Figure 1 which represents a plot of a five cluster data set with 10 points per cluster. The boundary limits for each cluster are defined by the dotted lines. Given the rather strong clustering present in the data, it would seem that any stopping rule which performs poorly would have rather questionable general applicability. It would be hard to believe that a method that fails on the present data would be able to perform better on less well-defined data structures.

Although the nature of the cluster generation process should have ensured that the clusters exhibited the desired properties, two external criterion measures also were used to verify the suitability of the data for purposes of testing the stopping rules. The criterion

measures use information external to the clustering process to validate recovery. In the present case, the external information was the knowledge of the true cluster structure in the data. The external criteria used were the Jaccard index (Milligan, 1981a) and the adjusted Rand statistic (Morey & Agresti, 1984). These two indices have been found to provide the best characteristics of those criteria in common use (Milligan, Soon, & Sokol, 1983; Milligan, 1983).

The results for both the Jaccard and the adjusted Rand indices indicated that distinct clustering was present at the correct level in the hierarchy in the vast majority of data set solutions. Out of a total of 432 solutions, the indices indicated optimal recovery at the correct level either 413 or 412 times. (In fact, cluster recovery was perfect at the correct level in 400 cases.) This information combined with the details concerning the cluster construction process indicates that the clustering in the test data was quite strong.

Stopping Rules

The stopping rules selected for study come from a wide variety of sources and disciplines. In addition to the statistical and psychometrics literature, such procedures have been published in the fields of pattern recognition, biological sciences, geology, and computing sciences, among others. The coverage is fairly extensive, but no claim for completeness is made. An attempt to include the techniques that have been incorporated in statistical packages was made. Techniques from a number of different approaches were included such as multivariate normal theory, graph theory, nonparametrics, etc. Finally, the best six internal criterion measures identified by Milligan (1981a) were examined in the present experiment.

It was considered desirable to examine only those procedures that were method independent. That is, the procedure should not be dependent on the usage of a given clustering method. Examples of method dependent procedures would be the techniques developed by Hall, Duda, Huffman, and Wolf (1973), Wong (1982), and Wong and Schaak (1982). Although only hierarchical methods were used to generate the clustering solutions in the present study, virtually all of the rules can be adopted for use with nonhierarchical procedures.

Similarly, the procedure should provide an automatic decision rule to eliminate the problems of human subjectivity. Thus, graphical methods requiring human judgment were not included in the study (see Andrews, 1972; Fleiss, Lawlor, Platman, & Fieve, 1971; Gengerelli, 1963; Gower, 1975; Jancey, 1966; Thorndike, 1953; among others). Likewise, indices with control parameters which have not been fully defined or developed were omitted (Rubin, 1967). Furthermore, those procedures requiring information external to the clustering process were omitted (Hubert & Baker, 1977; Hansen & Milligan, 1981). Finally, stopping rules applicable only under restricted conditions, such as one dimensional data sets, were not considered (Engleman & Hartigan, 1969; Hartigan, 1978).

The following discussion of the individual indices has been organized in terms of their final order of performance with the best indices appearing first. It should be recognized that the ordering must not be construed as valid for all possible data structures that might be encountered in real life research. Rather, the ordering reflects the performance of the rules for data sets where the clustering is especially strong. Furthermore, the numbers assigned to the rules are intended more for internal reference purposes rather than to indicate strict ordering.

In the study, each rule was allowed to adopt the most favorable conditions to optimize its performance. Where appropriate, the index was tested by taking both the maximum (or minimum) value across hierarchy levels, or by taking the maximum difference between successive levels as indicating the correct number of clusters in the data. The most favorable outcome was used to represent the results of the measure. In some situ-

ations where both positive and negative criterion values could occur, both raw and absolute value scores were used with the index. Similarly, when the literature was ambiguous, both covariance and sum of squares and cross products matrices were examined. If a given procedure required the specification of control parameters, a series of experiments were conducted across a reasonable range of parameter values to determine the best conditions for optimal recovery. When the control parameters corresponded to critical scores for test statistics, it was often found that a rather conservative alpha-level was required to give optimal performance. It would appear that this was a result of the repeated testing that was necessary as one searched the hierarchy.

Finally, the determination of the optimal level was limited to the last 25 levels of the hierarchical solution for several indices (26 or fewer clusters in the data). This was due to the fact that some indices exhibited distracting patterns when the data contain nearly as many clusters as elements in the data set. Such patterns often would dominate the results when searching for solutions where few clusters actually existed. By adopting this overall strategy, each index was permitted to exhibit its best performance. Since the data possessed distinct, error-free cluster structure, the failure of any given index in the present experiment would indicate that the criterion may have little to offer in an applied clustering situation. (Readers with an extensive familiarity of the clustering literature may want to skip (or skim) the remainder of the present section and continue reading with the results section. The remainder of the present section can be referenced as needed when studying the results.)

1. *Calinski and Harabasz.* The Calinski and Harabasz (1974) index is computed as $[\text{trace } B/(k-1)]/[\text{trace } W/(n-k)]$ where n and k are the total number of items and the number of clusters in the solution, respectively. The B and W terms are the between and pooled within cluster sum of squares and cross products matrices. The maximum hierarchy level was used to indicate the correct number of partitions in the data.

2. $Je(2)/Je(1)$. Duda and Hart (1973) proposed a ratio criterion where $Je(2)$ is the sum of squared errors within cluster when the data are partitioned into two clusters, and $Je(1)$ gives the squared errors when only one cluster is present. The hypothesis of one cluster is rejected if the ratio is smaller than a specified critical value. The critical value is computed from a formula given in their text and is a function of several terms including a standard normal score, the number of dimensions, and the sample size. Several values for the standard score were tested and the best results were obtained when the value was set to 3.20. It should be noted that this procedure is applied only to that subportion of the data involved in the cluster merger. Clustering was continued in the present study until the hypothesis was first rejected.

3. *C-Index.* The C-index was reviewed in Hubert and Levin (1976). It is computed as $[d_w - \min(d_w)]/[\max(d_w) - \min(d_w)]$ where d_w is the sum of the within cluster distances. The index was found to exhibit excellent recovery characteristics by Milligan (1981a). The minimum value across the hierarchy levels was used to indicate the optimal number of clusters.

4. *Gamma.* This index represents an adaptation of Goodman and Kruskal's Gamma statistic for use in a clustering situation (Baker & Hubert, 1975). The index is computed as $[s(+) - s(-)]/[s(+) + s(-)]$ where $s(+)$ represents the number of consistent comparisons involving between and within cluster distances, and $s(-)$ represents the number of inconsistent outcomes (Milligan, 1981a). Maximum values were taken to represent the correct hierarchy level.

5. *Beale.* Beale (1969) proposed the use of an F-ratio to test the hypothesis of the existence of $C2$ versus $C1$ clusters in the data ($C2 > C1$). The F-ratio compared the increase in the mean square deviation from the cluster centroids as one moved from $C2$ to $C1$ clusters against the mean square deviation when $C2$ clusters were present. Beale argued

that standard F-tables could be used with $p(C2 - C1)$ and $p(n - C2)$ degrees of freedom where p is the number of dimensions and n is the sample size. It was found that the .005 significance level gave the best results for the present data. At each level of the hierarchy in the present situation, a test was conducted for 2 versus 1 clusters in the data. The two clusters involved in the merger at the specific level in the hierarchy were used in the test. Clustering continued until the hypothesis of one cluster was rejected.

6. *Cubic Clustering Criterion.* The cubic clustering criterion is the test statistic provided by the SAS programming package (Ray, 1982; Sarle, 1983). The index is the product of two terms. The first term is the natural logarithm of $(1 - E(R^2))/(1 - R^2)$ where R^2 is the proportion of variance accounted for by the clusters and its expected value is determined under the assumption that the data have been sampled from a uniform distribution based on a hyperbox. The second term is $((np/2)^{.5})/((.001 + E(R^2))^{1.2})$, where p is an estimate of the dimensionality of the between cluster variation. The constant terms were chosen on the basis of extensive simulation results (Sarle, 1983). The maximum value across the hierarchy levels was used to indicate the optimal number of clusters in the data.

7. *Point-Biserial.* For this index, a point-biserial correlation is computed between corresponding entries in the original distance matrix and a matrix consisting of 0/1 entries that indicate whether or not two points are in the same cluster. The index was found to have desirable recovery properties by Milligan (1980, 1981a). The maximum value was used to suggest the optimal number of clusters in the data.

8. $G(+)$. The $G(+)$ index was reviewed by Rohlf (1974) and examined by Milligan (1981a). The formula is $[2s(-)]/[n_d(n_d - 1)]$ where $s(-)$ is defined as for the gamma index and n_d is the number of within cluster distances. Minimum values were used to determine the number of clusters in the data.

9. *Mojena.* The Mojena stopping rule is widely known and has been the subject of some limited validation research (Blashfield & Morey, 1980; Mojena, 1977). The rule parallels a one-tail confidence interval based on the fusion values at each level in the hierarchy. Computationally, one takes the average fusion value and adds to it a critical score times a measure of the standard error of the fusion values from the entire hierarchy. The first occurrence where a fusion value exceeds this confidence limit suggests that the previous hierarchy level was optimal. Given Mojena's own results, only Rule 1 was examined. Although Mojena suggested critical values in the range of 2.75 to 3.50, it was found that the best performance in the present study was obtained with a score of 1.25. This was determined by repeating the experiment with critical values in the range of 1.00 to 3.50 in steps of .25. Only the results with 1.25 are presented. At this optimal level, the correct recovery rate was 289. The method did tend to exhibit some insensitivity to the critical value. At 1.00, the correct recovery rate was 268. At 2.00, the rate was 273.

10. *Davies and Bouldin.* Davies and Bouldin (1979) provided a general framework for measures of cluster separation. The overall index is defined as the average of indices computed from each individual cluster. An individual cluster index is taken as the maximum pairwise comparison involving the cluster and the other partitions in the solution. Each pairwise comparison is computed as the ratio of the sum of the within cluster dispersions from the two partitions divided by a measure of the between cluster separation. In the present experiment, the within cluster dispersion was computed as the average within cluster distance. The distance between cluster centroids was used as the measure of between cluster separation. The minimum value across the hierarchy levels was used to indicate the number of clusters in the data.

11. *Stepsize.* The stepsize criterion dates to before the work of Johnson (1967) and Sokal and Sneath (1963). This rather simple criterion involves examining the difference in fusion values between hierarchy levels. A large difference would suggest that the data was

overclustered in the last merger. Thus, the maximum difference was taken as indicating the optimal number of clusters in the data.

12. *Likelihood Ratio.* Wolfe (1970) proposed a likelihood ratio criterion to test the hypothesis of k clusters against k-1 clusters. The procedure is modeled after the traditional Wilks' likelihood ratio criterion and is based on the assumption of multivariate normality. Everitt (1981) conducted a Monte Carlo analysis of Wolfe's procedure and found that Wolfe's original formula for the degrees of freedom for the test appeared to be valid only for cases where the sample size is about ten times larger than the number of dimensions. Everitt's results did indicate that Hartigan's (1977) suggested formula for the degrees of freedom is not correct. More importantly, Binder (1978) has shown that the test statistic is not asymptotically distributed as chi-square in the first place. Thus, the test procedure is at best an approximation.

For purposes of the present study, Wolfe's original formulation, including his formula for the degrees of freedom, were used. To maintain sufficient sample size, the assumption of common covariance matrices was made. Further, in order to maintain minimal sample sizes within clusters, only the last 10 hierarchy levels were tested. The level based on k clusters which corresponded to the first significant test result was assumed to indicate the optimal solution. It was found that the best recovery was obtained when the significance level was set to .01.

13. $|\log(p)|$. Gnanadesikan, Kettenring, and Landwher (1977) suggested a stopping rule based on $|\log(p)|$ where p is the p-value obtained from Hotelling's T^2 test. The two clusters involved in the merger at a given level defined the two groups for purposes of the T^2 test and the dimensions were used as the variables in the analysis. In the present experiment, maximum difference values were used to define the hierarchy level corresponding to the optimal number of clusters in the data.

14. *Sneath.* Sneath (1977) developed a method for testing the distinctness of clusters based on a measure of overlap. The numerator of the t_w statistic is the distance between the centroids of the two clusters under consideration for merger. The denominator of the index is a measure of the dispersion or overlap of the two clusters. As such, the test is not based on all of the data in the cluster analysis, but only on those items involved in the partitions undergoing merger. The statistic is compared to a critical score obtained from a noncentral t-distribution. The hypothesis of one cluster is rejected if t_w exceeds the critical score. In the present experiment, clustering was allowed to continue through the hierarchy until the hypothesis was rejected. A 5% significance level for the t-score and a 1% significance level for the extent of cluster overlap were found to give the best recovery.

15. *Frey and Van Groenewoud.* Frey and Van Groenewoud (1972) proposed a general stopping rule when introducing their k-method of clustering. The index is the ratio of difference scores from two successive levels in the hierarchy. The numerator is the difference between the average between cluster distances from each of the two hierarchy levels. The denominator is the difference between the mean within cluster distances from the two levels. The authors proposed using a ratio score of 1.00 to identify the correct cluster level. The ratios often varied above and below 1.00. The best results occurred when clustering was continued until the ratio fell below 1.00 for the last series of times. At this point, the cluster level before this series was taken as the optimal partition. If the ratio never fell below 1.00, a one cluster solution was assumed. This occurred in five instances.

16. Log (SSB/SSW). Hartigan (1975) proposed a statistic based on sum of squares as an index of cluster recovery. The values SSB and SSW are the sum of squared distances between and within groups, respectively. For the purposes of the present study, the maximum differences between hierarchy levels were taken as indicating the correct number of clusters in the data.

17. *Tau.* The tau index was reviewed by Rohlf (1974) and tested by Milligan (1981a). Computationally, the standard nonparametric tau correlation is computed between corresponding entries in two matrices. The first matrix contains the distances between items and the second 0/1 matrix indicates whether or not each pair of points are within the same cluster. The maximum value in the hierarchy sequence was taken as indicating the correct number of clusters.

18. $\bar{c}/k^{.5}$. Ratkowsky and Lance (1978; also see Hill, 1980) introduced a criterion for determining the optimal number of clusters based on $\bar{c}/k^{.5}$. The value for $\bar{c}$ is equal to the average of the ratios of $(SSB/SST)^{.5}$ obtained from each dimension in the data. The optimal number of groups is taken as the level where this criterion exhibits its maximum value.

19. $n \log (|T|/|W|)$. Among the measures studied by Scott and Symons (1971), it was proposed that one might use n Log $(|T|/|W|)$ where T and W are the total and within cluster sum of squares and cross products matrices. Of course, n is the number of elements in the data set. Except for multiplication by a constant, this is the same index as examined by Arnold (1979). In the present study, it was found that this form of the index produced much better results than $|T|/|W|$. The maximum difference between hierarchy levels was used to suggest the correct number of partitions.

20. $k^2|W|$. Marriot (1971) proposed the use of the term $k^2|W|$ where k is the number of clusters in the current solution. The maximum difference between successive levels was used to determine the best partition level.

21. *Bock.* Bock (1977) offered a testing procedure based on an approach involving both density estimation and nonparametric U-statistics. In particular, since the clusters in the present data approximate multivariate normality, Bock's procedure based on this assumption was used. The procedure considers those two clusters involved in the next merger in the hierarchy. Clustering continues until the test statistic exceeds the critical score and the partition before the last merger is used as the final solution. A critical score using a stepsize of .08 provided the best recovery.

22. *Ball and Hall.* Ball and Hall (1965) suggested that the average distance of the items to their respective cluster centroids could serve as a useful measure of the number of clusters in the data. In the present situation, the largest difference between levels was used to indicate the optimal solution.

23. *Trace Cov W.* Given the success of the Calinski and Harabasz (1971) index and the ambiguity in the definition of the matrices in several articles, it was decided to examine trace Cov W. This index represents the trace of the within clusters pooled covariance matrix. Again, maximum difference scores were used.

24. *Trace W.* The trace W criterion has been one of the most popular indices suggested for use in clustering context (Edwards & Cavalli-Sforza, 1965; Friedman & Rubin, 1967; Orloci 1967; see also Fukunaga & Koontz, 1970). Furthermore, it is useful to note that the error sum of squares criterion (based on within cluster distances) is the same index as trace W. To determine the number of clusters in the data, maximum difference scores were used since the criterion increases monotonically with solutions containing fewer clusters.

25. *Lingoes and Cooper.* Lingoes and Cooper (1971) introduced a nonmetric probability clustering criterion in their PEP-I clustering program. (A typographical error occurred in their article. In their equation (4), an n should appear before the product sign.) The logic behind the index is based upon graph theory arguments and directly provides a p-value for comparison to a specified Type I error rate value. Clustering is continued until the p-value is no longer significant. The optimal Type I error rate was found to be .001 in the present study.

26. *Trace* $W^{-1}B$. The maximum difference in values of trace $W^{-1}B$ criterion was

used to indicate the optimal number of clusters. It was found that recovery was better when covariance matrices were used instead of sums of squares and cross products matrices (total correct recovery in the latter case was only 7 data sets). The original index was proposed by Friedman and Rubin (1967) as a basis for a nonhierarchical clustering method.

27. *Generalized distance.* Day (1969) suggested assuming that the data were drawn frcm a mixture of two multivariate normal distributions. A maximum likelihood estimate of the generalized distance between the two clusters could then be computed. The assumption of common population covariance matrices was made in the present study. The maximum increase in the distance value between hierarchy levels was used to indicate the point of overclustering. Reasonable recovery performance was obtained only when the last ten hierarchy levels were examined. Inclusion of all hierarchy levels produced at best, five correct recoveries.

28. *McClain and Rao.* The McClain and Rao (1975) CLUSTISZ program employed a criterion which consisted of the ratio of two terms. The first term was the average within cluster distance divided by the number of within cluster distances. The denominator value was the average between cluster distance divided by the number of between cluster distances. The minimum value of the index was found to give the best recovery information. Good (1982) proposed a similar index which is effectively the reciprocal of the McClain and Rao measure.

29. *Mountford.* Mountford (1970) developed a testing procedure to determine whether two clusters should be merged or kept separate. The numerator of the test statistic is computed as the sum of the average within cluster distances minus the average distance between clusters. The denominator is a measure of within cluster variation. The index is based only on those points actually involved in the merger at a given level. If the cluster merger involved less than four points, the index was set to zero. (It should be noted that numerical errors are present in the first computational example given in Mountford's article.) Mountford did provide approximate critical values. Both the 1% and 5% levels were tested in the present study. However, these limits always resulted in solutions containing too many clusters. Rather, the best recovery was obtained by taking the maximum difference between levels as indicating the correct number of clusters in the data.

30. $|T|/|W|$. Friedman and Rubin (1967) proposed as an alternative clustering criterion the ratio of the determinant of the total sum of squares and cross products matrix to the determinant of the pooled within cluster matrix. In some instances, the reciprocal of the criterion has been suggested. However, the same decision as to the number of clusters is reached in either case. In the present study, the difference between levels was used. Covariance matrices also were tested with the present index but no improvement in recovery was found.

Results

A variety of analysis and presentation modes were examined for the data generated by the present study. These included several graphical displays and tabular summaries. It appeared that the clearest method for revealing the behavior of the stopping rules and presenting the results of the study was a simple tabular format. For a given stopping rule, the frequency of solutions consisting of too few and too many clusters are presented along with the number of correct determinations. The total frequency counts are presented and are further broken down for data sets containing 2, 3, 4, and 5 clusters. Dashed entries in the table correspond to situations where the index is either undefined or constant. This can occur when only one cluster is present in the solution.

TABLE 1

Frequency of Indication of Correct and Incorrect Stopping Levels For the First Five Stopping Rules

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
1. Calinski and Harabasz					
2 or Fewer	-	-	1	0	1
1 too Few	-	12	6	0	18
Correct Level	96	95	97	102	390
1 too Many	3	0	3	6	12
2 too Many	4	0	1	0	5
3 or More	5	1	0	0	6
2. Je(2)/Je(1)					
2 or Fewer	-	0	0	0	0
1 too Few	16	4	3	0	23
Correct Level	77	101	103	107	388
1 too Many	4	2	1	1	8
2 too Many	7	0	0	0	7
3 or More	4	1	1	0	6
3. C-Index					
2 or Fewer	-	-	4	0	4
1 too Few	-	5	12	12	29
Correct Level	71	89	91	96	347
1 too Many	2	7	1	0	10
2 too Many	2	0	0	0	2
3 or More	33	7	0	0	40
4. Gamma					
2 or Fewer	-	-	8	0	8
1 too Few	-	9	16	12	37
Correct Level	74	86	83	96	339
1 too Many	3	7	1	0	11
2 too Many	2	0	0	0	2
3 or More	29	6	0	0	35
5. Beale					
2 or Fewer	-	3	4	0	7
1 too Few	34	8	5	8	55
Correct Level	57	87	95	92	331
1 too Many	0	1	0	0	1
2 too Many	4	0	0	0	4
3 or More	13	9	4	8	34

The results for the first five stopping rules are presented in Table 1. The entries for the line labeled "2 or Fewer" gives the number of occurrences where the stopping rule produced solutions with two or fewer clusters than were actually present in the data. That is, suppose that there were five clusters present in the data. If the stopping rule suggested

that there were either one, two, or three clusters present, then the result would be recorded in this category. The line labeled "1 too Few" indicates the number of solutions where the rule selected a level with one less cluster than present in the data. Similarly, the lines beginning with "1 too Many" or "2 too Many" specify the number of solutions where the rule indicated one or two clusters more than were actually present in the data, respectively. Finally, "3 or More" gives the frequency count for the number of solutions where the stopping rule selected a level containing 3 or more clusters than present in the data. For example, if the data possessed two clusters and the rule suggested that there were five or more clusters, the result would be recorded here.

Certainly, the results presented in Table 1 seem to indicate that stopping rules do exist which can be effective at determining the correct number of clusters in data which possess distinct clustering. The Jaccard and adjusted Rand external criterion measures noted earlier provide upper limits on the performance of a stopping rule. Given the fact that these external criteria produced 413 and 412 correct determinations, respectively, it would appear that the Calinski and Harabasz index and the $Je(2)/Je(1)$ rule developed by Duda and Hart (1973) provided excellent recovery. Further, when errors did occur, they tended to be near misses. The Duda and Hart index did have some difficulty at the level of two clusters where several solutions with too few clusters were found. Although recovery at the level of 2 clusters is not poor, it should be recognized that this stopping rule allows the user to test that only one cluster is present in the data. The Calinski and Harabasz index, on the other hand, performed rather consistently across the varying number of clusters. The Duda and Hart index sets a pattern which is seen in the results of the remaining three rules in Table 1. The C-index, Gamma, and Beale indices tended to follow the pattern set by the Duda and Hart procedure with a drop in recovery at the level of two clusters. In particular, the C-index and Gamma rules produced a substantial number of far misses at this level.

The results for the next five rules appear in Table 2. It is reassuring to find that the index used in the SAS package, the cubic clustering criterion, performed at a competitive rate. The criterion exhibited the highest rate of determining too many clusters seen so far, but it did produce a relatively low number of solutions with too few clusters. Thus, if the index makes a decision error, it is more likely to result in an incomplete clustering of the data. On the other hand, the point-biserial measure produced the lowest error rate for determining too many clusters. However, of the best ten indices it exhibited the highest error rate for too few clusters, even if most of these errors can be considered near misses. This pattern is counter to the one found with the external criteria and the other indices in Tables 1 and 2. As such, the index would appear to be less attractive. Finally, it is useful to note that of the ten indices found in Tables 1 and 2, eight exhibited their lowest recovery rates for solutions where two clusters were present. It would seem that the two cluster case is the most difficult structure for the stopping rules to detect.

The pattern in Table 2 for the Mojena (1977) rule is consistent with Mojena's own published results where better recovery was found when three to five clusters were present in the data. When analyzing the index for the present study, it became apparent to the authors that the critical value needed for optimal recovery at two clusters was different from the value required for optimal performance when four or five clusters were present. Of course, such adjustments of critical values in an applied context are not possible since this presupposes knowledge of the true number of clusters in the data. Before leaving Table 2, one should note the large number of far misses for the Mojena, $G(+)$, and Davies and Bouldin indices for data sets with two clusters present.

The indices in Table 3 represent somewhat more mediocre stopping rules and completes the first half of the 30 indices. Again, the three patterns seen in Tables 1 and 2 can be detected in this set. The stepsize and $|\log(p)|$ criteria produced their best recovery at

TABLE 2

Frequency of Indication of Correct and Incorrect Stopping Levels for the Rules Numbered 6 Through 10

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
6. Cubic Clustering Criterion					
2 or Fewer	-	0	0	0	0
1 too Few	13	5	4	0	22
Correct Level	67	88	82	84	321
1 too Many	5	7	10	13	35
2 too Many	6	0	7	9	22
3 or More	17	8	5	2	32
7. Point-Biserial					
2 or Fewer	-	-	13	12	25
1 too Few	-	23	29	31	83
Correct Level	94	83	66	65	308
1 too Many	8	1	0	0	9
2 too Many	4	0	0	0	4
3 or More	2	1	0	0	3
8. $\underline{G}(+)$					
2 or Fewer	-	-	4	0	4
1 too Few	-	8	16	12	36
Correct Level	52	70	79	96	297
1 too Many	1	4	1	0	6
2 too Many	0	0	0	0	0
3 or More	55	26	8	0	89
9. Mojena					
2 or Fewer	-	0	1	2	3
1 too Few	0	2	13	14	29
Correct Level	20	84	93	92	289
1 too Many	12	12	1	0	25
2 too Many	19	8	0	0	27
3 or More	57	2	0	0	59
10. Davies & Bouldin					
2 or Fewer	-	-	3	0	3
1 too Few	-	15	18	9	42
Correct Level	54	72	72	89	287
1 too Many	2	4	3	0	9
2 too Many	5	1	3	0	9
3 or More	47	16	9	10	82

the level of two clusters. This pattern was first seen with the point-biserial rule. Also consistent with the point-biserial results, these two indices produced relatively few solutions which suggested that too many clusters were present. The likelihood ratio rule generated fairly constant recovery rates across the number of clusters factor in a manner

TABLE 3

Frequency of Indication of Correct and Incorrect Stopping Levels for the Rules Numbered 11 Through 15

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
11. Stepsize					
2 or Fewer	-	-	37	29	66
1 too Few	-	51	18	11	80
Correct Level	96	56	53	68	273
1 too Many	6	1	0	0	7
2 too Many	1	0	0	0	1
3 or More	5	0	0	0	5
12. Likelihood Ratio					
2 or Fewer	-	0	0	0	0
1 too Few	12	0	4	0	16
Correct Level	64	72	64	68	268
1 too Many	12	16	17	25	70
2 too Many	9	10	16	9	44
3 or More	11	10	7	6	34
13. $\lvert\log(p)\rvert$					
2 or Fewer	-	-	43	52	95
1 too Few	-	35	19	11	65
Correct Level	78	71	45	43	237
1 too Many	11	1	1	1	14
2 too Many	10	0	0	1	11
3 or More	9	1	0	0	10
14. Sneath					
2 or Fewer	-	28	12	2	42
1 too Few	54	12	17	14	97
Correct Level	34	51	66	83	234
1 too Many	0	1	1	0	2
2 too Many	0	1	2	1	4
3 or More	20	15	10	8	53
15. Frey & Van Groenewoud					
2 or Fewer	-	1	2	7	10
1 too Few	0	0	20	18	38
Correct Level	0	76	79	77	232
1 too Many	1	3	3	6	13
2 too Many	20	2	0	0	22
3 or More	87	26	4	0	117

similar to the Calinski and Harabasz rule. It is interesting to note that the stepsize rule corresponds to the simplest criterion in the experiment whereas the likelihood ratio procedure is one of the most computationally complex with an underlying theoretical development. Yet, there is little difference in the performance between these two rules. The

remaining indices in Table 3 produced better recovery when four or five clusters were present, a pattern first set by the Duda and Hart rule.

Finally, with respect to Tables 2 and 3, none of the methods displayed any clear balance between the two types of error rates. The point-biserial, stepsize, $|\log(p)|$, and Sneath rules indicated solutions with too few clusters at a rate 2 to 11 times greater than solutions with too many clusters. This was counter to the trend established by the external criteria. The remaining methods, the cubic clustering criterion, $G(+)$, Mojena, Davies and Bouldin, likelihood ratio, and Frey and Van Groenewoud, tended to choose solutions with too many clusters. When solutions with too many clusters occurred for some indices, such as for the Davies and Bouldin procedure, the errors would often occur rather high in the hierarchy (indicating far too many clusters in the data).

Unfortunately, no general statement of cause as to the reason behind the distinct performance patterns can be offered by the present authors. This is due to the fact that the underlying formulations of the indices within groups are so fundamentally different and similarities between the sets of rules can be found.

Starting with Table 4, one now has procedures which produced as many or more errors than correct determinations. Beginning with the Tau index, all remaining stopping rules have an error rate for either too few or too many clusters that by itself exceeds the correct recovery rate. Furthermore, a new performance pattern begins to appear with this table. Several stopping rules exhibited a tendency to specify a constant number of clusters in the data. For example, the Log (*SSB/SSW*) rule displayed a bias to indicate the three cluster solution regardless of the true number of clusters in the data. The pattern can be detected by noting the very high oorrect recovery rate at the level of three clusters, the high error rate for solutions with too many clusters for the two cluster data sets, and the inflated rates for the other kind of error with the four and five cluster data sets. Similar patterns can be seen in the $n \log(|T|/|W|)$ and $k^2|W|$ rules. The Tau and $\bar{c}/k^{.5}$ indices follow the pattern first set by the point-biserial index. However, both of these rules experienced a considerable amount of difficulty in recovery when five clusters were present in the data. Finally, all of the indices in Table 4, except log (*SSB/SSW*), produced a notable bias to generate more solutions with too few clusters as opposed to solutions with too many clusters.

Table 5 includes results for one of the most widely proposed stopping rules. The relatively poor performance of Trace W is informative in that stopping rules which have intuitive statistical appeal may, in fact, have little validity. This rule is identical to the error sum of squares criterion and has been a rather popular index for use in the clustering context, including its use in determining the number of clusters in the data. The results for the index would suggest that it has been an especially unfortunate choice given the available alternatives. With respect to the other methods in Table 5, it is possible that the Bock procedure might perform better for data sets involving fewer dimensions (three or less). (A similar comment can be made for the Log (*SSB/SSW*) and the cubic clustering criterion.)

Table 6 presents the results for the last five stopping rules. When considering the indices, it is worth noting that a recovery rate corresponding to the random selection of hierarchy levels would be about 9. Thus, Mountford's (1970) rule appears to have been functioning at this level. In studying this index, it became clear to the present authors that the asymptotic critical values provided by Mountford were too small in magnitude. This resulted in a vastly inflated rate for solutions with too many clusters. Similarly, the recovery rates for the generalized distance and McClain and Rao indices functioned only slightly better than chance level. Furthermore, the $|T|/|W|$ index never detected the correct number of clusters in the data in 432 attempts. It is clear from the table that except for the Trace $W^{-1}B$ criterion, the indices were selecting solutions that suggested far too many clusters were present.

TABLE 4

Frequency of Indication of Correct and Incorrect Stopping Levels for the Rules Numbered 16 Through 20

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
16. Log(SSB/SSW)					
2 or Fewer	-	-	0	22	22
1 too Few	-	0	66	20	86
Correct Level	0	104	42	66	212
1 too Many	52	3	0	0	55
2 too Many	19	0	0	0	19
3 or More	37	1	0	0	38
17. Tau					
2 or Fewer	-	-	25	52	77
1 too Few	-	28	52	46	126
Correct Level	85	77	30	10	202
1 too Many	7	2	1	0	10
2 too Many	7	0	0	0	7
3 or More	9	1	0	0	10
18. $\bar{c}/k^{.5}$					
2 or Fewer	-	0	23	52	75
1 too Few	0	27	60	49	136
Correct Level	88	80	25	7	200
1 too Many	12	1	0	0	13
2 too Many	3	0	0	0	3
3 or More	5	0	0	0	5
19. nLog(|T|/|W|)					
2 or Fewer	-	-	0	62	62
1 too Few	-	0	74	33	107
Correct Level	0	104	32	13	149
1 too Many	32	2	0	0	34
2 too Many	21	0	0	0	21
3 or More	55	2	2	0	59
20. $k^2\|W\|$					
2 or Fewer	-	-	0	73	73
1 too Few	-	0	92	6	98
Correct Level	0	104	15	27	146
1 too Many	95	4	1	2	102
2 too Many	8	0	0	0	8
3 or More	5	0	0	0	5

Discussion

The results of the present study are interesting due to the rather wide range of performance which was found for the stopping rules. It appears that a relatively accurate set of rules has been identified that may be of help to applied researchers in determining the

TABLE 5

Frequency of Indication of Correct and Incorrect Stopping Levels for the Rules Numbered 21 Through 25

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
21. Bock					
2 or Fewer	-	0	18	30	48
1 too Few	0	63	34	38	135
Correct Level	74	15	31	22	142
1 too Many	17	9	8	7	41
2 too Many	5	7	6	1	19
3 or More	12	14	11	10	47
22. Ball & Hall					
2 or Fewer	-	-	0	63	63
1 too Few	-	0	85	44	129
Correct Level	0	104	23	1	128
1 too Many	56	3	0	0	59
2 too Many	18	0	0	0	18
3 or More	34	1	0	0	35
23. Trace Cov ($\underline{W}$)					
2 or Fewer	-	-	0	81	81
1 too Few	-	0	91	27	118
Correct Level	0	104	17	0	121
1 too Many	59	3	0	0	62
2 too Many	18	0	0	0	18
3 or More	31	1	0	0	32
24. Trace $\underline{W}$					
2 or Fewer	-	-	0	88	88
1 too Few	-	0	92	20	112
Correct Level	0	104	16	0	120
1 too Many	62	3	0	0	65
2 too Many	19	0	0	0	19
3 or More	27	1	0	0	28
25. Lingoes and Cooper					
2 or Fewer	-	0	55	45	100
1 too Few	0	47	18	27	92
Correct Level	37	30	17	16	100
1 too Many	24	6	1	1	32
2 too Many	7	7	3	1	18
3 or More	40	18	14	18	90

number of clusters in a data set. Furthermore, other stopping rules were found to be particularly ineffective in producing correct determinations in data sets that possessed especially strong and distinct cluster structure. Among the best six internal criteria found by Milligan (1981a), four placed in the best ten in the present study (*C*-index, gamma,

TABLE 6

Frequency of Indication of Correct and Incorrect Stopping Levels for the Rules Numbered 26 Through 30

Stopping Rule	Number of True Clusters				
	2	3	4	5	Overall
26. Trace $\underline{W}^{-1}\underline{B}$					
2 or Fewer	-	0	0	69	69
1 too Few	0	0	56	19	75
Correct Level	0	52	23	9	84
1 too Many	85	48	27	3	163
2 too Many	15	7	0	0	22
3 or More	8	1	2	8	19
27. Generalized Distance					
2 or Fewer	-	12	17	27	56
1 too Few	16	0	17	11	44
Correct Level	5	22	11	9	47
1 too Many	6	4	1	7	18
2 too Many	8	1	5	7	21
3 or More	73	69	57	47	246
28. McClain & Rao					
Too Few	-	0	0	0	0
Correct Level	9	5	5	6	25
1 too Many	2	2	1	1	6
2 too Many	1	0	0	0	1
3 or More	96	101	102	101	400
29. Mountford					
Too Few	0	0	0	0	0
Correct Level	1	6	1	2	10
1 too Many	0	0	7	4	11
2 too Many	3	3	3	3	12
3 or More	104	99	97	99	399
30. $\lvert\underline{T}\rvert/\lvert\underline{W}\rvert$					
Too Few	-	0	0	0	0
Correct Level	0	0	0	0	0
3 or More	108	108	108	108	432

point-biserial, and $G(+)$), one placed in the middle third (tau), and one was found in the lower third (McClain and Rao). Thus, there is some tendency for the better internal criterion measures also to perform among the best stopping rules, but the consistency in performance is not guaranteed.

Although the results concerning the best methods in Table 1 are encouraging, it should be noted that the findings are likely to be somewhat data dependent. It would not surprise the present authors to find that the ordering of the indices would change if different data structures were used. However, it would seem rather unlikely that an index

found among the upper third of the indices in the present study would place in the lowest third if different data structures were used.

As an example of data dependence, the Duda and Hart (1973) procedure requires the computation of a critical value for decision purposes. One of the entries used in the computation is described as a standard normal score. In the present study, a value of 3.20 was found to produce the best recovery. Values of, say 3.10 and 3.30, resulted in slightly fewer correct determinations. It would not seem unwarranted to anticipate that the optimal value might change if the data structure changed. Hence, the optimal critical score is likely to be data dependent. Further research on the effect of alternative data structures on the control parameters is needed. It is useful to note that control parameters are not necessary to develop an effective stopping rule. The Calinski and Harabasz procedure, the best stopping rule found in the present experiment, is not dependent on the specification of a critical score.

On the other hand, it is remarkable to note that the Duda and Hart procedure is computed only from the information provided by the items involved in the last cluster merger. Thus, the effective sample size is markedly reduced from the overall sample size for most decisions made with the procedure. Thus, given the results of the present study, it would appear that the effective power of the rule is quite reasonable. Similar comments also hold for the Beale criterion. It is reassuring to discover techniques which are not dependent on especially large sample sizes.

Certainly, one can argue that other rules may be able to show improved performance under conditions which differ from the present study. For example, the likelihood ratio rule (12) would be anticipated to perform better for data sets consisting of a much larger sample size. Although this would be nice to show, it misses the point. If the better performing methods from the present study also perform well in the presence of larger sample sizes, then there is little reason to pursue the asymptotic methods since they do not provide consistently superior performance across the useful sample size range. Similarly, consistent performance across alternative data structures would be desirable. If one or more consistent rules could not be found, then future research should result in recommendations that are data dependent. As long as these dependencies can determined from the data, such results should benefit the applied user.

Some comments concerning the use of difference scores are in order. A difference score is a comparison of criterion values from one hierarchy level to the next. In fact, several stopping rules incorporate comparisons between levels in their basic definition (rules 2, 5, 11, 12, and 15). Other more subjective proposals have included the process of plotting a criterion value, such as trace W, against the number of clusters in the solution (Jancey, 1966). A marked drop in the criterion value followed by a sequence of small differences might indicate an optimal number of clusters. This process is analogous to the case where one attempts to find the proper number of dimensions from a plot of Kruskal's stress in multidimensional scaling. Difference scores were adopted for some stopping rules in the present study because they optimized the performance of certain criteria (rules 13, 16, 19, 20, 22, 23, 24, 26, 27, 29, 30). It is worth noting that most of these applications gave relatively poor recovery of the correct number of clusters. As such, this application of difference scores may be of little value. As one reviewer to the present paper noted, there is no clear theoretical justification for the use of difference scores for many of the criteria.

A number of authors have advocated the development of rigorous statistical techniques for cluster analytic situations (Fleiss & Zubin, 1969; Goodall, 1966). The progress in this area has been slow, no doubt due in part to the immense distributional problems. Mathematical statisticians have tended to begin the analysis by assuming multivariate normality (see, for example, Cohen, 1967; Day, 1969; Hartigan, 1975; Jain & Waller,

1978; Lee, 1979; Naus, 1966; among others). This practice has been seriously questioned by Gower (1981) who argued that better alternative statistical strategies may be available. Certainly, the rather poor performance of the standard multivariate normal criteria in the present study, such as trace W, trace $W^{-1}B$, and $|T|/|W|$, lend credence to Gower's arguments. An examination of the references in this area indicates that a majority of the approaches based on multivariate normality were published in the more traditional statistical journals, while other approaches tended to appear in more applied sources. The field of clustering might be well served if mathematical statisticians considered alternative distributional strategies.

The results of the present study represent the conclusion of a three-part research project. The first segment (Milligan, 1980) addressed the issue of whether a given clustering method could find the true underlying cluster structure in error-free and error-perturbed data. This phase assumed that the researcher knew the correct number of clusters in the data. The second phase of the project (Milligan, 1981a) determined whether a given internal criterion measure could discriminate between recovery of true cluster structure and arbitrary partitions of random data. The results of the second phase would allow a researcher to construct a reasonably valid test for the existence of cluster structure. This is accomplished by using a simulation procedure to provide the appropriate sampling distribution (Milligan & Sokol, 1980). Again, this process assumed the prior knowledge of the true number of clusters in the data. The present work, the third in the sequence, is the first that directly addressed the question as to how many clusters are actually present in a data set. Although the results of the three-part project are dependent on the nature of the simulated data, the validation process appears to be sound and can provide a framework for extending our knowledge of clustering procedures to other data structures.

References

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

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

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: Stanford Research Institute. (NTIS No. AD 699616).

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

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika, 65*, 31–38.

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. (1977). On tests concerning the existence of a classification. In *First international symposium on data analysis and informatics* (Vol. 2, pp. 449–464). Rocquencourt, France: IRIA.

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

Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics, 9*, 15–28.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*, 224–227.

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

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

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

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

Englemann, L., & Hartigan, J. A. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association, 64*, 1647–1648.

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

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

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.

Frey, T., & Van Groenewoud, H. (1972). A cluster analysis of the D-squared matrix of white spruce stands in Saskatchewan based on the maximum–minimum principle. *Journal of Ecology, 60*, 873–886.

Fukunaga, K., & Koontz, W. L. G. (1970). A criterion and an algorithm for grouping data. *IEEE Transactions on Computers, C-19*, 917–923.

Gengerelli, J. A. (1963). A method for detecting subgroups in a population and specifying their membership list. *Journal of Psychology, 5*, 457–468.

Gnanadesikan, R., Kettenring, J. R., & Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses. *Bulletin of the International Statistical Institute, 47*, 451–463.

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.

Goodall, D. W. (1966). Hypothesis testing in classification. *Nature, 221*, 329–330.

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.

Gower, J. C. (1981, June). *Is classification statistical*? Paper presented at the meeting of the Classification Society, Toronto.

Hall, D. J., Duda, R. O., Huffman, D. A., & Wolf, E. E. (1973). *Development of new pattern recognition methods*. Los Angeles: Aerospace Research Laboratories. (NTIS No. AD 7726141).

Hansen, R. A., & Milligan, G. W. (1981). Objective assessment of cluster analysis output: Theoretical considerations and empirical findings. *Proceedings of the American Institute for Decision Sciences*, 314–316.

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*, New York: Academic Press.

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

Hill, R. S. (1980). A stopping rule for partitioning dendrograms. *Botanical Gazette, 141*, 321–324.

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

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83*, 1072–1080.

Jain, A. K., & Waller, W. G. (1978). On the number of features in the classification of multivariate gaussian data. *Pattern Recognition, 10*, 365–374.

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

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

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

Lingoes, J. C., & Cooper, T. (1971). PEP-I: A FORTRAN IV (G) program for Guttman-Lingoes nonmetric probability clustering. *Behaviorial Science, 16*, 259–261.

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

McClain, J. O., & Rao, V. R. (1975), CLUSTISZ: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research, 12*, 456–460.

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. (1981c, June). *A discussion of procedures for determining the number of clusters in a data set*. Paper presented at the meeting of the Classification Society, Toronto.

Milligan, G. W. (1983). Characteristics of four external criterion measures. In J. Felsenstein, (Ed.), *Proceedings of the 1982 NATO Advanced Studies Institute on Numerical Taxonomy* (pp. 167–173). New York: Springer-Verlag.

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

Milligan, G. W., Soon, S. C., & Sokol, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*, 40–47.

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

Morey, L., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement, 44*, 33–37.

Mountford, M. D. (1970). A test for the difference between clusters. In G. P. Patil, E. C. Pielou, & W. E. Waters (Eds.), *Statistical Ecology* (Vol. 3, pp. 237–257). University Park, Pa.: Pennsylvania State University Press.

Naus, J. I. (1966), A power comparison of two tests of non-random clustering. *Technometrics, 8*, 493–517.

Orloci, L. (1967). An agglomerative method for classification of plant communities. *Journal of Ecology, 55*, 193–206.

Perruchet, C. (1983). *Les épreuves de classifiabilité en analyses des données* [Statistical tests of classifiability] (Tech. Rep. NT/PAA/ATR/MTI/810). Issy-Les-Moulineaux, France: C.N.E.T.

Ray, A. A. (Ed.). (1982). *SAS user's guide: Statistics.* Cary, North Carolina: SAS Institute.

Ratkowsky, D. A., & Lance, G. N. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal, 10*, 115–117.

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

Rubin, J. (1967). Optimal classification into groups: An approach for solving the taxonomy problem. *Journal of Theoretical Biology, 15*, 103–144.

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

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

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., & Sokal, R. R. (1973). *Numerical taxonomy.* San Francisco: Freeman.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy.* San Francisco: Freeman.

Thorndike, R. L. (1953). Who belongs in a family? *Psychometrika, 18*, 267–276.

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

Manuscript received 9/16/83
Revision received 5/17/84
Final version received 10/10/84