PSYCHOMETRIKA—VOL. 46, NO. 2.
JUNE, 1981

# A MONTE CARLO STUDY OF THIRTY INTERNAL CRITERION MEASURES FOR CLUSTER ANALYSIS

GLENN W. MILLIGAN

THE OHIO STATE UNIVERSITY

A Monte Carlo evaluation of thirty internal criterion measures for cluster analysis was conducted. Artificial data sets were constructed with clusters which exhibited the properties of internal cohesion and external isolation. The data sets were analyzed by four hierarchical clustering methods. The resulting values of the internal criteria were compared with two external criterion indices which determined the degree of recovery of correct cluster structure by the algorithms. The results indicated that a subset of internal criterion measures could be identified which appear to be valid indices of correct cluster recovery. Indices from this subset could form the basis of a permutation test for the existence of cluster structure or a clustering algorithm.

Key words: classification, numerical taxonomy, permutation tests.

An inherent problem in the use of a clustering algorithm in practice is the difficulty of validating the resulting data partition. This is a particularly serious issue since virtually any clustering algorithm will produce partitions for any data set, even random noise data which contain no cluster structure. Thus, an applied researcher is often left in a quandary as to whether the obtained clustering of a real life data set actually represents significant cluster structure or an arbitrary partition of random data. Hence, some type of statistic which reflects the degree of the recovery of true cluster structure would be clearly desirable.

Two types of criterion measures can be used to assess the significance of a data partition [Sneath, 1969]. An external criterion index uses information which is obtained from outside the clustering process in order to evaluate the resulting data partition. Unfortunately, in an applied research setting, valid external criterion statistics are not usually available since the true cluster structure is not known on an a priori basis. A related version of external analysis involves using the obtained cluster partitions to define the treatment levels in an ANOVA context. A variable which was not used to cluster the data is used as the dependent measure. However, the validity of the external criterion variable in an applied situation is usually open to question. Failure to find significant group differences may be due to either a lack of cluster structure or an invalid criterion variable.

On the other hand, an internal criterion measure uses information obtained strictly from within the clustering process. Internal criterion measures typically reflect the goodness-of-fit between the input data and the resulting data partition. Although several reviews of internal criteria are available in the literature [Cormack, 1971; Jardine & Sibson, 1971; Rohlf, 1974], critical comparisons of the recovery characteristics of the indices have not been conducted. It seems that the trend in the clustering literature has been for authors to continue to introduce new statistics while providing little comparative information. The present report was designed to determine if any internal criterion could indicate whether a given partition of the data recovered a significant portion of the true cluster structure or whether any structure exists at all.

Four hierarchical algorithms were used to cluster simulated data and included the complete link (furthest neighbor), single link (nearest neighbor), group average (UPGMA),

Requests for reprints should be sent to Glenn W. Milligan, Faculty of Management Sciences, 356 Hagerty Hall, The Ohio State University, Columbus, Ohio 43210.

0033-3123/0600-3000$00.75/0

and Ward's minimum variance method. An analysis of the four algorithms with respect to recovery of the true cluster structure for the distance based data used in the present study can be found in Milligan [1980]. The four algorithms were found to vary in their ability to recover the cluster structure, particularly for the error-perturbed data sets. Although recovery was adversely affected by the error conditions, the methods still produced recovery rates which were significantly greater than chance levels.

In the present study, both the external and internal criteria were computed at the hierarchy level which corresponded to the exact number of clusters known to be present in the data. It should be noted that the algorithms were used essentially to generate a variety of typical partition solutions. It seems reasonable that the results of the study can be generalized to the analysis of data partitions produced by other hierarchical and nonhierarchical procedures which generate nonoverlapping solutions.

In terms of an outline for the present paper, the method section discusses the generation of the artificial data and the statistics which were examined. The construction of the cluster structure in the data is discussed first and is followed by a description of the error-perturbation process. The method section concludes with a discussion of the various internal and external statistics. The result section first evaluates the performance of internal criteria with respect to the external indices. A subset is identified which consists of the best internal measures. The result section concludes by examining the superior subset in detail. The discussion section considers several applications and limitations of the results of the study.

## *Method*

### *Data Sets*

A total of 108 error-free data sets were generated which consisted of fifty points each. A two-dimensional plot of one of the data sets is presented in Figure 1 as an example. Three replications were taken from a 36-cell design created by three factors which controlled the overall characteristics of the clusters. The first design factor specified whether 2, 3, 4, or 5 clusters would be present in the data. The second design factor involved embedding the clusters in either a 4, 6, or 8 dimensional space. The third factor consisted of three different patterns for the assignment of the points to the clusters. The first level of the third factor distributed the points as equally as possible across the clusters. The second and third levels required that one cluster must always contain 10% or 60% of the points respectively, with the rest distributed equally across the remaining clusters.

The clusters were defined in a euclidean space and all clusters possessed convex boundaries. In order to satisfy the requirement of external isolation, cluster overlap was not permitted on the first dimension of the variable space. That is, the clusters were required to occupy disjoint regions of space. The separation between cluster boundaries was computed as:

$$f \times (S_i + S_j) \tag{1}$$

where $f$ is the separation factor and $S_i$ and $S_j$ are the standard deviations for clusters $i$ and $j$, respectively. The value of $f$ was randomly selected from the uniform interval of .25 to .75. (The random number generator used in the study was the LLRANDOM package, see Learmonth and Lewis, Note 5. See Dudewicz, 1976, for information regarding the quality of the uniform generator and Dudewicz, Note 2, for information concerning the normal generator.)

Cluster overlap was permitted on the remaining dimensions of the space. Any pair of clusters may or may not have overlapped on any of the remaining dimensions. The plot in Figure 1 represents two dimensions on which overlap was permitted. The actual location of

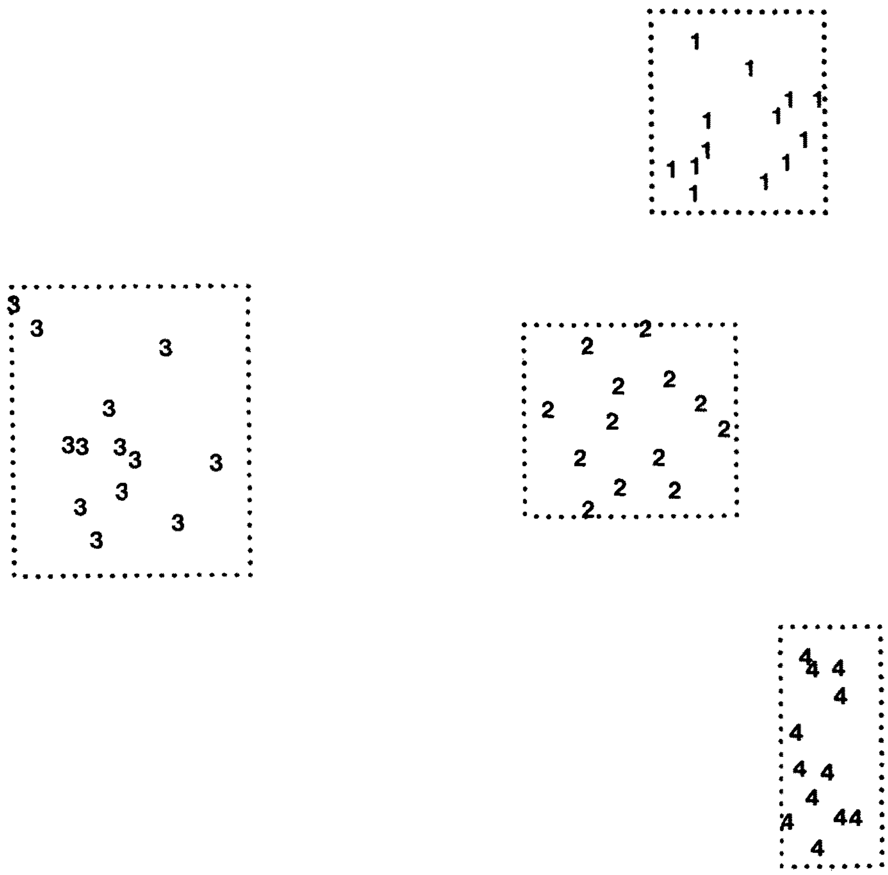

FIGURE 1
Example plot of a four cluster data set

the cluster boundaries was chosen randomly and without regard to the location of the other clusters present. Thus, a variety of cluster configurations was possible.

The boundary length of a cluster on each dimension was selected randomly from the uniform interval of 10 to 40 units. The boundary length was defined to be three standard deviations for the cluster. As shown in Figure 1, a cluster typically did not have the same standard deviation on any two dimensions and two clusters were not likely to have the same standard deviation on any given dimension.

The actual location of the points within cluster boundaries was determined with the aid of a truncated multivariate normal random number generator. The centroid of the multivariate distribution was located at the geometric center of the cluster. The variance-covariance matrix was defined as a diagonal matrix. A truncated distribution ensured that all points fell within the cluster boundaries. Any generated point with coordinate values which exceeded 1.5 standard deviations from the mean on any dimension was rejected and another location was examined.

This method of cluster construction generated a point distribution which was greatest near the geometric center of the cluster. The point density gradually decreased as the distance from the centroid increased. As with any Monte Carlo experiment, several arbitrary decisions had to be made. However, it was felt that this approach generated clusters which possessed intuitive appeal. As can be seen in Figure 1, the clusters did exhibit the properties of external isolation and internal cohesion as defined by Cormack [1971].

*Error-Perturbation Process*

A total of 432 data sets were used to study the internal criterion measures. Each of the following error conditions produced a total of 108 data sets.

*1. Error-free data sets.* The error-free data sets which were generated directly from the cluster construction program were analyzed by each of the four clustering algorithms. This condition produced data sets which provided substantial recovery of the true cluster structure by the algorithms [Milligan, 1980].

*2. Error perturbation of the distances.* The true distances in the Condition 1 data sets were error perturbed to form a new distance matrix. This condition corresponds to the situation where a true cluster structure exists but a noisy measurement process is involved. Specifically, the interpoint distances were recomputed as:

$$\sum(d_{ij} - d_{ik} - e_{ijk})^2. \tag{2}$$

The $d_{ij}$ entry represents the coordinate value for point $j$ on dimension $i$. The $e_{ijk}$ values were obtained from a univariate normal random number generator with a mean of zero. The standard deviation was obtained by taking the sum of the stardard deviations on dimension $i$ for the two clusters which contained points $j$ and $k$.

*3. Addition of random noise dimensions.* Two random noise dimensions were added to the existing set of dimensions which defined the true cluster structure in the error-free data sets. The random noise dimensions possessed a range equal to the length of the dimension on which cluster overlap was not permitted. A uniform random generator was used to determine the location of each point on the dimensions.

*4. Random noise data sets.* In order to provide baseline rates for the criterion measures and values representative of data which does not contain cluster structure, 108 random noise data sets were generated. Properties such as the number of dimensions, the number of hypothesized clusters and their relative sizes, and the range on each dimension matched the characteristics of one of the error-free data sets. A uniform distribution generator was used to locate the points on each dimension of the space.

*External Criteria*

The Rand [1971] and Jaccard [Anderberg, 1973] statistics served as measures of true cluster recovery. The indices are very general in that they reflect the degree of correct recovery for a given partition level. Another index, the kappa statistic, also has been used as an external criterion measure [Blashfield, 1976; Edelbrock, 1979]. However, both Edelbrock and McLaughlin [Note 3] and Milligan and Isaac [1980] found that the Rand and kappa statistics correlate above .975. Hence, the two indices are essentially identical for practical applications.

The computation of the Rand and Jaccard statistics can be based on a simple classification scheme as shown in Table 1. Frequency counts for each of the four cells are tabulated from the $n(n - 1)/2$ pairs of points in a data set. For example, cell $a$ represents the number of pairs of points in a data set which were classified together in both the criterion definition of cluster structure and the obtained algorithm solution. The Rand statistic is computed as $(a + b)/(a + b + c + d)$ and is identical to the simple matching coefficient where the numerator represents the number of consistent outcomes and the denominator gives the total number of comparisons. The Jaccard index, on the other hand, ignores the count for different cluster consistent matches ($d$) in both the numerator and denominator: $a/(a + b + c)$. The Jaccard index was included because some researchers have questioned the use of $d$ in the computation of the Rand statistic [Downton & Brennan, Note 1; Fowlkes & Mallows, Note 4]. However, the data in the present study indicated that the correlation between the Rand and Jaccard indices was .937. Hence, the issue of the count for different cluster matches may not pose serious problems in the evaluation of cluster

Table 1

Pairwise Classification Scheme for the External Criteria

| Algorithm Solution | Criterion Solution | |
|---|---|---|
| | Pair in Same Cluster | Pair Not In Same Cluster |
| Pair in Same Cluster | $\underline{a}$ | $\underline{b}$ |
| Pair Not in Same Cluster | $\underline{c}$ | $\underline{d}$ |

Note. $\underline{a} + \underline{b} + \underline{c} + \underline{d} = \underline{n}(\underline{n} - 1)/2$.

recovery. Fowlkes and Mallows have also noted that the variance of the Rand statistic is somewhat restricted. The results of the present study produced a standard deviation of .20 for the Rand statistic and .32 for the Jaccard index. Apparently, the variance restriction problem is less severe for the Jaccard statistic.

*Internal Criterion Measures*

Thirty internal criterion measures were examined in the present study. A listing of the indices with an appropriate reference for each is given in Table 2. Although this does not exhaust the set of indices which have been proposed, it does provide a fairly extensive coverage over the various approaches to the formulation of internal criterion measures. Several indices such as the Gamma, $C$ Index, and Tau coefficients require only nominal or ordinal scale data. On the other hand, the Point-Biserial, $W/B$, and most of the Jardine and Sibson measures require interval or ratio scale information. Some indices are analogous to test statistics used in MANOVA (Trace $W^{-1}B$, $|W|/|T|$), or use a maximum likelihood formulation. Others are based on a variety of simple counting principles (Gamma, Tau, $G(+)$, $G(-)$). Finally, some indices were orginally derived for entirely different purposes. The Gamma, Point-Biserial, and Tau coefficients represent various types of correlation measures which have been adopted for use in clustering. Likewise, most of the indices reviewed by Hubert and Levin [1976] were first proposed as performance measures for use in studying categorical clustering in free recall. However, the indices do compare a partition of data elements with a matrix of similarities between elements. Hence, the indices could be used as internal criterion measures in the more general clustering context.

## *Results*

Logically, if a given internal criterion is succeeding in indicating the degree of correct cluster recovery, the index should exhibit a close association with an external criterion which reflects the actual degree of recovery of structure in the proposed partition solution. Thus, in order to determine the validity of the internal criteria, each measure was correlated with each external criterion. The correlations were based on data from all methods and error conditions. All four error conditions were included in the data base in order to

Table 2

Pearson and Spearman Correlations Between the External Criteria and Internal Criteria

| Internal Criterion | Reference | Pearson | | Spearman | |
|---|---|---|---|---|---|
| | | Rand | Jaccard | Rand | Jaccard |
| Gamma | Baker and Hubert [1975] | .91 | .77 | .89 | .82 |
| $\underline{C}$ Index | Hubert and Levin [1976] | .90 | .75 | .89 | .82 |
| Point-Biserial | Milligan [1980] | .89 | .78 | .88 | .82 |
| Tau | Rohlf [1974] | .87 | .74 | .84 | .78 |
| $\underline{W}/\underline{B}$ | McClain and Rao [1975] | .82 | .74 | .85 | .73 |
| $\underline{G}(+)$ | Rohlf [1974] | .78 | .73 | .82 | .81 |
| Tau $\underline{A}$ | Hubert and Levin [1976] | .79 | .68 | .77 | .72 |
| Tau $\underline{C}$ | Hubert and Levin [1976] | .79 | .68 | .77 | .72 |
| Somer's Gamma | Hubert and Levin [1976] | .79 | .68 | .77 | .72 |
| Modified Ratio of Repetition | Hubert and Levin [1976] | .75 | .58 | .76 | .58 |
| Adjusted Ratio of Clustering | Hubert and Levin [1976] | .72 | .50 | .73 | .51 |
| $\underline{\Delta}_0^*$ | Jardine and Sibson [1971] | .70 | .47 | .77 | .51 |
| Fagan's Index | Hubert and Levin [1976] | .70 | .45 | .70 | .47 |
| $\underline{\Delta}_{1/2}^*$ | Jardine and Sibson [1971] | .67 | .37 | .71 | .50 |
| Alpha | Jardine and Sibson [1971] | .67 | .37 | .71 | .50 |
| Deviation Index | Hubert and Levin [1976] | .65 | .38 | .65 | .40 |
| $\underline{Z}$-Score Index | Hubert and Levin [1976] | .63 | .31 | .68 | .43 |
| Ratio of Repetition | Hubert and Levin [1976] | .63 | .38 | .64 | .42 |
| Pi | Jardine and Sibson [1971] | .60 | .62 | .74 | .71 |
| Trace $\underline{W}^{-1}\underline{B}$ | Friedman and Rubin [1967] | .60 | .45 | .69 | .58 |
| $\underline{G}(-)$ | Rohlf [1974] | .60 | .48 | .56 | .53 |
| $\underline{\Delta}_1^*$ | Jardine and Sibson [1971] | .58 | .25 | .61 | .34 |
| Log Likelihood | Hartigan [1975] | .49 | .25 | .74 | .22 |
| $\underline{D}$ Index | Hubert and Levin [1976] | .45 | .20 | .70 | .48 |
| $\lvert\underline{W}\rvert / \lvert\underline{T}\rvert$ | Friedman and Rubin [1967] | .38 | .27 | .52 | .49 |
| $\hat{\Delta}_{1/2}$ | Jardine and Sibson [1971] | .24 | .07 | .28 | .12 |
| $\Sigma\ \underline{d}_i / \underline{D}$ | Williams et al., [1971] | .22 | .05 | .25 | .02 |
| $\hat{\Delta}_1$ | Jardine and Sibson [1971] | .20 | .02 | .23 | .06 |
| Trace $\underline{W}$ | Friedman and Rubin [1967] | .18 | .49 | .18 | .52 |
| $\hat{\Delta}_0$ | Jardine and Sibson [1971] | .06 | .03 | .09 | .06 |

Note: Several indices have inverse relationships with the Rand and Jaccard statistics. For purposes of clarity, only the absolute value of the correlations are reported. Large sample 99% nonsimultaneous confidence interval width for the Pearson and Spearman correlation coefficient is .062 ($\underline{n}$ = 1728).

provide recovery values which were obtained from a variety of cluster structures ranging from error-free to error-perturbed and random noise data. All four algorithms were included in order to provide a variety of solution partitions for each underlying structure. Both the Pearson and Spearman correlations are reported in Table 2. The Spearman correlation was computed in the event that a somewhat nonlinear relationship existed between a given internal criterion and the Rand or Jaccard statistic. The results demonstrate that the relative performance of the internal criteria was quite varied.

When compared to the Rand statistic, the best five internal indices for either the Pearson or Spearman correlation seem to be the Gamma, *C* Index, Point-Biserial, Tau, and *W*/*B* statistics. The Jaccard criterion indicates that these five criteria should be considered plus the *G*(+) index. The set of six indices was chosen because pairwise tests between the best internal criterion and each successive index resulted in acceptance of the hypothesis that the correlations were equal for either the Pearson or Spearman measure. The test where one variable is common to each correlation was used [see Guilford & Fruchter, 1973, p. 167]. The overall Type I error rate was limited to .05 with the use of the Bonferroni approach for each external criterion.

Mean recovery values for the two external and six internal criteria by each error condition are presented in Table 3. Mean recovery values reflect the degree of recovery of true cluster structure by the algorithms for an external criterion and the suggested degree of recovery for an internal criterion. An examination of the values produced by the Rand and Jaccard statistics indicated that the clustering algorithms had little difficulty in recovering the true cluster structure in the error-free data sets (Condition 1). The algorithms had varying degrees of success in recovering the cluster structure in Error Condition 2 (distance perturbation condition) and Condition 3 (addition of random noise dimensions). The values for Condition 4 (random noise data) represent baseline rates for the recovery measures with data which do not possess distinct cluster structure. Repeated measures ANOVA

Table 3

Mean Recovery Values by Error Condition for the External Criteria and the Best Six Criterion Measures

| Criterion | Error Condition | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Rand (.14)[a] | .99 | .89 | .85 | .53 |
| Jaccard (.18) | .98 | .86 | .70 | .28 |
| Gamma (.14) | .98 | .88 | .78 | .44 |
| C Index (.07) | .01 | .06 | .11 | .27 |
| Point-Biserial (.14) | .85 | .71 | .63 | .33 |
| Tau (.11) | .63 | .55 | .50 | .27 |
| W/B (.12) | .29 | .44 | .61 | .81 |
| G(+) (.03) | .01 | .02 | .05 | .10 |

Note: The C Index, W/B, and G(+) have inverse relationships with the Rand and Jaccard statistics.

[a] Parenthetical entries indicate the average within cell standard deviation.

for each external criterion indicated highly significant differences between error conditions, even when the conservative degrees of freedom correction was used. Post hoc tests between cell means with the Newman-Kuels procedure indicated that all pairwise comparisons within each row were significant at the .05 level for the two external criteria. Thus, as would be expected, cluster recovery from the error-free data was better than recovery with the error-perturbed data. Further, the recovery values in Conditions 1, 2, and 3 were each significantly greater than the baseline values for Condition 4. (All ANOVA test procedures used in the study were conducted with and without appropriate variance stabilizing transformations. In all cases, identical results with respect to inference were obtained.)

When considering the recovery means across error conditions for the six internal criterion measures, essentially identical results were obtained. In every case, the Newman-Keuls procedure indicated that all pairwise differences within each row were significant. Thus, each internal criterion succeeded in differentiating between data which contained cluster structure and data consisting of random noise. The measures could even differentiate between error-free data and error perturbed cluster structure.

In order to determine whether the effectiveness of the internal criteria was dependent on the type of clustering algorithm, Pearson correlations between the external criteria and the internal measures were computed for each of the four algorithms used in the study. The correlations reported in Table 4 are fairly stable despite the differences in the characteristics of the clustering algorithms. None of the pairwise differences between correlations for the same internal and external criterion across methods are statistically reliable. Thus, the performance of the internal criteria does not seem to be dependent on the type of clustering algorithm.

Overall, the six internal criterion measures appear to be valid indices of true cluster recovery. The criteria apparently can distinguish between random noise data and data which contain distinct clustering. To some extent, the indices seem to be somewhat substitutable since their pairwise correlations are quite substantial as seen in Table 5. The Tau, *W/B*, and *G*( + ) indices may be somewhat less desirable in this respect.

A factorial analysis of variance using the data obtained for each of the six internal criteria as dependent variables was conducted for the three design factors used to generate

Table 4

Pearson Correlations Between the External Criteria and the Best Six Internal Criteria by Clustering Method

| Criterion | Single Link | | Complete Link | | Group Average | | Ward's Method | |
|---|---|---|---|---|---|---|---|---|
| | Rand | Jaccard | Rand | Jaccard | Rand | Jaccard | Rand | Jaccard |
| Gamma | .89 | .85 | .93 | .80 | .90 | .79 | .92 | .78 |
| C Index | -.88 | -.82 | -.92 | -.81 | -.90 | -.79 | -.93 | -.79 |
| Point-Biserial | .88 | .81 | .91 | .85 | .87 | .79 | .91 | .85 |
| Tau | .86 | .78 | .89 | .83 | .84 | .75 | .89 | .83 |
| W/B | -.84 | -.81 | -.84 | -.77 | -.80 | -.75 | -.82 | -.73 |
| G(+) | -.75 | -.81 | -.88 | -.71 | -.85 | -.76 | -.86 | -.67 |

Note: Large sample 99% nonsimultaneous confidence interval width for the difference between two correlations is .175 (n = 432).

Table 5

Pearson and Spearman Correlations
Between the Best Six Internal Criteria

| Criterion | Gamma | C Index | Point-Biserial | Tau | W/B |
|---|---|---|---|---|---|
| C Index | -.99(-.99) | | | | |
| Point-Biserial | .96 (.95) | -.97(-.96) | | | |
| Tau | .93 (.88) | -.95(-.90) | .97 (.95) | | |
| W/B | -.91(-.95) | .91 (.96) | -.88(-.89) | -.83(-.82) | |
| G(+) | -.89(-.96) | .84 (.93) | -.78(-.87) | -.74(-.79) | .83 (.91) |

Note: The parenthetical entries are the Spearman correlations (n = 1728).

the cluster structure. The analyses indicated that the only consistent significant effect occurred for the factor involving the number of clusters. The average recovery values for the two external and six internal indices for this factor are presented in Table 6. The results for the Rand statistic are consistent with the findings of Fowlkes and Mallows [Note 4]. The authors indicated that the Rand statistic approaches 1.00 as the number of clusters increases. Apparently, the Jaccard index exhibits a reverse trend. All of the internal criteria seem to possess a strong tendency to increase or decrease as the number of clusters increases. This indicates that the comparison of values between hierarchy levels may not be justified in an applied analysis.

In order to determine whether the number of clusters affects the relationship between the external and internal criteria, the correlations between the criteria were recomputed at each level of the number of clusters. No systematic trend was found for the best five internal criteria and significance test results did not produce rejection rates at the .05 or .01 levels greater than that which would be expected by chance. Only the $G(+)$ index appeared to be significantly affected by this factor. Lower correlations were observed for three cluster data sets than for two, four, or five cluster sets.

Table 6

Mean Recovery Values by the Number of Clusters
for the External Criteria and the Best Six Internal Criterion Measures

| Criterion | Number of Clusters | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| Rand (.20)[a] | .80 | .80 | .83 | .85 |
| Jaccard (.32) | .76 | .70 | .68 | .68 |
| Gamma (.24) | .68 | .76 | .81 | .83 |
| C Index (.12) | .18 | .11 | .09 | .08 |
| Point-Biserial (.23) | .54 | .64 | .67 | .67 |
| Tau (.17) | .42 | .51 | .52 | .50 |
| W/B (.21) | .65 | .55 | .50 | .45 |
| G(+) (.05) | .06 | .05 | .04 | .03 |

[a] Parenthetical entries indicate the average within cell standard deviation.

### *Discussion*

A subset of internal criterion measures have been identified in the present study which appear to be valid indices of cluster recovery and have been found to be superior to other proposed criteria. It should be noted that both metric and monotone invariant indices are included in this subset. Thus, an appropriate index is available for either ordinal or interval scale data. Slight changes in the superior subset could occur if different external criterion statistics are considered. However, it is expected that the subset would be fairly stable since all external criterion indices should reflect the same degree of relative cluster recovery. Thus, the external criteria would exhibit high intercorrelations and should be more or less substitutable for each other.

The computational formulas for the best six internal criteria are presented in Table 7. For the Gamma, Tau, and $G(+)$ indices, $s(+)$ represents the number of times where two points not clustered together had a larger distance than two points which were in the same cluster, whereas $s(-)$ represents the reverse outcome. For the $C$ Index, Point-Biserial, and $W/B$ indices, $d_w$ represents the sum of within cluster distances, $d_b$ is the sum of between cluster distances, and $\bar{d}_w$ and $\bar{d}_b$ represent the respective means. The absolute minimum and maximum values for $d_w$ given the data must also be used for the $C$ Index. Finally, $s_d$ is the standard deviation of all distances and $t$ is the number of comparisons of two pairs of points where both pairs represent within cluster comparisons or both pairs are between cluster comparisons.

Several applications of the results are possible. First, an internal criterion index could be used to form the basis of a permutation procedure to test for the existence of cluster structure. Sneath [1969] has argued that a random distribution of the data in the variable space seems to be the most generally useful null hypothesis. An applied researcher could generate an approximate sampling distribution for the index with this approach and compare the observed criterion value against the distribution. The approach was adopted by Rohlf and Fisher [1968] in the development of a hypothesis test for the cophenetic correlation coefficient (a measure of the recovery of the entire hierarchy structure). However,

Table 7

Computational Formulas for the Best Six Internal Criteria

| Criterion | Formula |
|---|---|
| Gamma | $[s(+) - s(-)] / [s(+) + s(-)]$ |
| $C$ Index | $[d_w - \min(d_w)] / [\max(d_w) - \min(d_w)]$ |
| Point-Biserial | $[\bar{d}_b - \bar{d}_w] [f_w f_b / n_d^{\,2}]^{\frac{1}{2}} / s_d$ |
| Tau | $[s(+) - s(-)] / [(n_d(n_d - 1)/2 - t)(n_d(n_d - )/2)]^{\frac{1}{2}}$ |
| $W/B$ | $[d_w/f_w] / [d_b/f_b]$ |
| $G(+)$ | $[2s(-)] / [n_d(n_d - 1)]$ |

Note: The total number of distances is represented by $n_d$ while $f_w$ and $f_b$ indicate the number of within cluster and between cluster distances, respectively.

some care must be taken in the specification of the process used to generate the sampling distribution. Virtually any clustering algorithm will tend to maximize the internal criteria, even for random noise data. Thus, the criterion values which are used to generate the distribution should be obtained from cluster solutions of random noise data as in the Rohlf and Fisher approach. This fact was overlooked by McClain and Rao [1975] in the development of a permutation test procedure for their cluster statistic. Milligan and Mahajan [1980] found that when the significance level of the McClain and Rao test was set to .05, the observed Type I error rate was as high as 1.00.

The interpretation as to what constitutes "random noise" data is open to question. Milligan and Mahajan used a uniform distribution to construct random data. Rohlf and Fisher [1968] and Arnold [1979] used both a uniform and a multivariate normal distribution. The consequences of the selection of different null distributions for determining the location of the data units in the variable space has not been examined in detail for the internal criteria listed in Table 2.

It is important to note that other testing procedures have been proposed [Lingoes & Cooper, 1971; Sneath 1979a, 1979b]. In particular, the Lingoes and Cooper procedure provides an exact statistical test for the existence of clusters for either nonmetric or interval data. Unfortunately, these procedures came to the attention of the author after the present study had been completed.

A different application of the results would involve the development of a clustering algorithm. Several authors have developed algorithms which explicitly attempt to maximize the value of a goodness-of-fit statistic. The three indices in Table 2 proposed by Friedman and Rubin [1967] were offered as optimization criteria in their clustering algorithm. Unfortunately, none of the three indices ranked in the upper third of the total set for either external criterion. On the other hand, Johnson [1967] suggested than an optimal clustering algorithm would be one which maximized the rank order correlation between the input data and the resulting hierarchical solution. Given that the Gamma and Tau coefficients were found to be among the best internal criteria, Johnson's speculation seems correct. An algorithm which is designed to optimize one or more of the top ranking internal criterion measures found in the present study would have a basis for validity which is lacking for many existing clustering algorithms.

A third application was suggested by a reviewer. One of the internal criteria could be computed for each of two hierarchical trees, presumably based on the same stimuli or even the same data set. The resulting values of the criterion for one tree could then be plotted against the value for the other tree at each level of the hierarchy. In this manner, the relationship between the trees could be studied and possibly an optimal hierarchy level could be identified.

The generality of the present results are, of course, limited due to the Monte Carlo nature of the analysis. For example, the variables were uncorrelated within clusters and euclidean distance was used as the dissimilarity measure. Differing conceptualizations of cluster structure or differing spatial metrics may alter the results. Certainly, most if not all of the indices in Table 2 would not work well if the clusters were nonconvex, substantially overlapped, or if fairly elongated clusters were present in the data. Likewise, some indices may fail if the data is embedded in a different type of metric space. Only limited information is available with respect to the last issue. Milligan [1980] found that the observed correlation between the Rand statistic and the point-biserial index was actually slightly higher when the data were generated in an ultrametric space as opposed to a euclidean space. Thus, differing spatial metrics may have little impact on the effectiveness of the internal criterion measures.

A particularly troublesome issue in clustering is the discrimination between two fairly similar solutions. For example, a researcher may need assistance in determining how many clusters are present in a data set. Such a situation occurs with the use of a hierarchical clustering algorithm where an applied researcher usually wants to select a specific partition level as the final solution. Procedures for determining which hierarchy level is the best representation of the data have not been well developed. This is partially due to the fact that the discrimination between two similar solutions can place an extreme performance demand on any recovery measure. Consider two solution partitions for the same data where one solution is exactly correct with respect to the underlying cluster structure. Let the second partition be identical to the first except for the misassignment of a single data unit. A statistic which could reliably discriminate between these two solutions would have to be extremely powerful. Further, the results of the present study indicate that the comparison of values between hierarchy levels for the best internal criteria may not be justified. Hence, it appears that a solution to this problem is likely to be a difficult task.

## REFERENCE NOTES

1. Downton, M., & Brennan, T. *Comparing classifications: An evaluation of several coefficients of partition agreement.* Paper presented at the meeting of the Classification Society, Boulder, Colorado, June 1980.
2. Dudewicz, E. J. *IRCCRAND-The Ohio State University random number generator package* (Tech. Rep. No. 104). Columbus, Ohio: The Ohio State University, Department of Statistics, 1974.
3. Edelbrock, C., & McLaughlin, B. *Intraclass correlations as metrics for hierarchical cluster analysis: Parametric comparisons using the mixture model.* Paper presented at the meeting of the Classification Society, Gainesville, Florida, April 1979.
4. Fowlkes, E. B., & Mallows, C. L. *A new measure of similarity between two hierarchical clusterings and its use in studying hierarchical clustering methods.* Paper presented at the meeting of the Classification Society, Boulder, Colorado, June 1980.
5. Learmonth, G. P., & Lewis, P. A. W. *Naval Postgraduate School random number generator package LLRANDOM* (Tech. Rep. NPS55LW73061A). Monterey, Calif.: Naval Postgraduate School, Department of Operations Research and Administrative Sciences, 1973.

## REFERENCES

Arnold, S. J. A test for clusters. *Journal of Marketing Research*, 1979, *16*, 545–551.

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

Baker, F. B., & Hubert, L. J. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 1972, *70*, 31–38.

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

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

Dudewicz, E. J. Speed and quality of random numbers for simulation. *Journal of Quality Technology*, 1976, *8*, 171–178.

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

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

Guilford, J. P., & Fruchter, B. *Fundamental statistics in Psychology and Education.* New York: McGraw-Hill, 1973.

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

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

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

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

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

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

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

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

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

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

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

Rohlf, F. J., & Fisher, D. R. Tests for hierarchical structure in random data sets. *Systematic Zoology*, 1968, *17*, 407–412.

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

Sneath, P. H. A. Basic program for a significance test for clusters in UPGMA dendrograms obtained from squared euclidean distance. *Computer Geosciences*, 1979, *5*, 127–137.(a)

Sneath, P. H. A. Basic program for a significance test for 2 clusters in euclidean space as measured by their overlap. *Computer Geosciences*, 1979, *5*, 143–155.(b)

Williams, W. T., Clifford, H. T., & Lance, G. N. Group-size dependence: A rationale for choice between numerical classifications. *Computer Journal*, 1971, *14*, 157–162.

*Manuscript received 7/15/80*
*Final version received 2/6/81*