Measuring the Power of Hierarchical Cluster Analysis

FRANK B. BAKER and LAWRENCE J. HUBERT*

The concept of power for monotone invariant clustering procedures is developed via the possible partitions of objects at each iteration level in the obtained hierarchy. At a given level, the probability of rejecting the randomness hypothesis is obtained empirically for the possible types of partitions of the *n* objects employed. The results indicate that the power of a particular hierarchical clustering procedure is a function of the type of partition. The additional problem of estimating a "true" partition at a certain level of a hierarchy is discussed briefly.

1. INTRODUCTION

During the last decade, many researchers in the behavioral and biological sciences have been concerned with developing and utilizing techniques of data analysis that can be subsumed under the general term of "clustering procedures." Most of this work has involved the development of strategies for clustering and, to a lesser extent, providing a theoretical context for the many suggested techniques. With some exceptions (Hubert [9], Ling [12]), the theoretical justification of a particular approach to clustering has not been imbedded in any standard statistical framework. Even in the most recent literature, e.g., Sneath and Sokal [13] or Anderberg [1], the emphasis is still on the intuitive reasonableness of a technique and the practical details of implementation rather than on testing hypotheses that relate to the existence of clusters.

If the field is to advance beyond simple data description, a much firmer connection is needed between the extremely useful and informative clustering schemes now available and statistical theory. Specifically, it should be possible to test hypotheses pertaining to certain characteristics of the structures yielded by a particular clustering procedure and to obtain the power of these tests against certain types of alternatives.

The major intent of this article is to discuss a statistical concept of power for those hierarchical clustering schemes that depend solely on the ordinal information contained in the proximity measures between the elements of a basic object set. Several aspects of power within the clustering context are then investigated in detail using one specific example and two well-known hierarchical clustering techniques, commonly called the single-link and complete-link methods.

2. BACKGROUND

Prior to examining the problem of hypothesis testing and statistical power within the clustering framework, it would be valuable to present a brief review of what hierarchical clustering procedures attempt to do. Suppose S is a set of n objects $\{O_1, O_2, \cdots, O_n\}$ and $s(\cdot,\cdot)$ is a positive real-valued symmetric function on $S \times S$. Increasing index values are assigned to object pairs by the function $s(\cdot,\cdot)$ as a way of representing increasing dissimilarity and, in addition, it is assumed for the present that distinct values of $s(\cdot,\cdot)$ are given to distinct unordered object pairs. If $L(S)$ denote the set of all partitions of S, then a *hierarchical clustering procedure* can be defined as any technique that uses the function $s(\cdot,\cdot)$ to construct a partition hierarchy $(\ell_0, \ell_1 \cdots \ell_{n-1})$ satisfying the following properties:

1) $\ell_k \in L(S)$, $0 \leq k \leq n-1$.
2) ℓ_0 is the trivial partition containing all objects in separate classes; ℓ_{n-1} is the trivial partition defined by a single all-inclusive class.
3) ℓ_k is a proper refinement of ℓ_{k+1} in the sense that ℓ_{k+1} is obtained from ℓ_k by uniting two of the object classes in ℓ_k.

If the construction of the particular hierarchy depends solely on the rank order of the object pairs as defined by $s(\cdot,\cdot)$, then the clustering procedure is referred to as *monotone invariant.*

The two monotone invariant hierarchical clustering procedures employed in the present study are usually called the single-link and complete-link methods. Both of these techniques can be defined by varying the criterion imposed to unite two subsets in the partition ℓ_k to form ℓ_{k+1}. In particular, if ℓ_k consists of the subsets $L_1, \cdots, L_{n-k}$, then L_i and L_j are united to form ℓ_k if

$$\text{Single-link: } \min\{s(O_i, O_j) \mid O_i \in L_i, O_j \in L_j\} = \min\{\min\{s(O_r, O_t) \mid O_r \in L_r, O_t \in L_t\}\}\ , \quad 1 \leq r \neq t \leq n-k\ ;$$

$$\text{Complete-link: } \max\{s(O_i, O_j) \mid O_i \in L_i, O_j \in L_j\} = \min\{\max\{s(O_r, O_t) \mid O_r \in L_r, O_t \in L_t\}\}\ , \quad 1 \leq r \neq t \leq n-k\ .$$

These definitions can be operationalized in the following manner. The inter-object proximity measures,

[*] Frank B. Baker is professor and Lawrence J. Hubert is associate professor, both with the Department of Educational Psychology, University of Wisconsin, Madison, Wis. 53706. The authors are listed alphabetically.

© Journal of the American Statistical Association
March 1975, Volume 70, Number 349
Applications Section

$s(O_i, O_j)$, are rank ordered from smallest to largest and scanned sequentially from rank 1 to rank $n(n-1)/2$. Using the single-link criterion, two groups in the partition ℓ_k will be merged to form a new group whenever a proximity value for an object pair is reached corresponding to one object from each group. Thus, a single link between subsets containing O_i and O_j results in the merger of two subsets in ℓ_k. Under the complete-link alternative, two existing groups at ℓ_k are merged when all possible object pairs in the union have been encountered in the scan. Thus, all possible links must be found before two groups can be merged into a single subset in ℓ_{k+1}.

These two strategies are generally considered the prototypes for all monotone invariant hierarchical clustering schemes; the reader is referred to [13] for a more comprehensive presentation.

At a very general level, formulating the notion of power within the clustering context can be developed along the lines employed in the standard randomization tests of hypotheses. Following Ling's [12] argument, the null hypothesis under test is that the $n(n-1)/2$ proximity values have been assigned at random to the object pairs. If randomness cannot be rejected, the hierarchy yielded by the clustering procedure will be considered an artifact of the method rather than the result of any structure inherent in the data. To test such a hypothesis under the general randomization framework developed by Fisher [6] (see also [10, 3, 12]), the obtained proximity values and a specific clustering procedure are used to construct a partition hierarchy along with the calculation of some statistic G, defined as a function on the partition hierarchy. The basal proximity measures are then permuted over the $n(n-1)/2$ object pairs, analyzed by the same clustering procedure, and the value of the statistic G recalculated.

Since this process is repeated for all possible permutations of the proximity values, the empirical distribution of G is obtained. A rejection region can be defined on this distribution; and if the value of G obtained from the basal data falls within this region, the hypothesis of randomness is rejected where the size of the rejection region, α, is the probability of making a Type I error. In using randomization techniques for analysis of variance, the statistic G is usually chosen to be sensitive to differences in location.

Using this approach, Collier and Baker [4] investigated power within the ANOVA context in terms of the standard permutation procedures; in particular, they added known nonnullities to the basal data after randomization and, consequently, constructed the empirical distribution of the F statistic under an alternative hypothesis. Following this paradigm, they were able to obtain the power of the permutation F test for a number of different alternatives.

Within the clustering context, however, a randomization approach to power involves new difficulties. First, the alternatives of interest may correspond to complete partition hierarchies rather than a value of a parameter or parameters; consequently, the investigator has a much more difficult task in picking a specific alternative of interest. In addition, the number of partition hierarchies or possible alternative hypotheses is a function of the number of objects clustered, and for large samples, the number of possible alternatives is exceedingly large. Second, even if an alternative hierarchy were specified, imposing some type of nonnullity on the basal data after randomization is somewhat more complex conceptually. Finally, the best choice of a statistic G is not immediately obvious since we may select a statistic that depends on the complete hierarchy or, alternatively, one that is a function of only certain of the partitions within the hierarchy. In short, the development of a concept of power for clustering depends on some reasonable means for specifying the alternatives against which power is desired as well as the choice of some statistic G that is sensitive to these alternatives.

One of the prime uses of the concept of power from a statistical point of view is in comparing various hypothesis testing procedures and providing a basis for selecting the best strategy for a given situation. The same concern is relevant in clustering where specific hierarchical clustering procedures could be differentially sensitive to given alternatives, where sensitivity is expressed through some notion of power related to a particular statistic G.

3. CONSIDERATIONS OF POWER WITHIN THE CLUSTERING FRAMEWORK

A reasonable procedure for specifying a "null" hypothesis of randomness is available for clustering but any development of the corresponding concept of power demands a categorization of the alternative hypotheses of interest. Hopefully, the approach taken here is consistent with problems that are important for researchers who use clustering procedures for data analysis.

Ideally, the choice of a clustering procedure should be based on sensitivity to a particular partition hierarchy as the alternative of interest, but most investigators would probably find this task impossible. Nevertheless, it may be reasonable to specify an alternative of importance by the number of objects within each cluster at some specific partition level of the hierarchy. By reducing the specifications of the alternative hypothesis from a complete hierarchy to a particular partition within that hierarchy, an investigation of power can be made somewhat more tractable and possibly some recommendations developed on what clustering procedures should be used.

The choice of a single type of partition as the "true" configuration underlying the proximity measures deserves further explanation since hierarchical clustering schemes are generally viewed as imposing a much more comprehensive structure on the data. First, it would be possible to define a "true" structure by means of a complete partition hierarchy; in fact, some aspects of this problem have already been attacked by Baker [2] and Hubert [9]. The present approach is somewhat broader in the sense that if a "true" hierarchy is to be detected, then as a necessary prerequisite, the "true" partitions at each level must be approximated adequately as well.

Second, it is considerably easier to define the exhaustive set of all possible structures using the notion of a "true" partition rather than a "true" hierarchy. In fact, a "true" structure defined by a single partition is really equivalent to a class of "true" hierarchies where each of the hierarchies must pass through the chosen type of partition at the appropriate level. Consequently, if rather definite power comparisons can be made at each level of a partition hierarchy using some statistic G, then a logical extension of these individual results should provide an immediate recommendation for the detection of complete hierarchies. An analogous approach is taken in the development of clique structures in sociograms in which the proximity measures are thresholded at a particular level to define a simple linear graph.

A third reason for conceptualizing alternative hypotheses around a particular partition level is derived from the manner in which cluster analysis is commonly applied. Many users of cluster analysis are really not interested in the complete hierarchy but only in the construction of single partitions that organize the proximity data in some substantively interpretable fashion. For instance, although any hierarchical clustering procedure will produce a complete sequence of partitions, many applied researchers usually limit major substantive interpretations to one particular level m.

In summary, the authors believe that an investigation of power against an alternative hypothesis defined as a particular partition rather than as a complete hierarchy is a necessary step in the development of a more general statistical framework for clustering. More important for our present state of knowledge, such an approach also makes the overall problem reasonably tractable.

One possible measure of partition adequacy, formulated by Hubert [8] and called α_ℓ, may be given the following precise interpretation: suppose $T_\ell(O_{i'}, O_{j'}) = 0$ if $O_{i'}$ and $O_{j'}$ belong to the same subset in ℓ and 1 otherwise, and define

$$d_\ell(O_i, O_j) = |\{\{O_r, O_t\} : T_\ell(O_r, O_t) = 1 \text{ and } s(O_r, O_t) < s(O_i, O_j)\}| \, , \tag{3.1}$$

where $|\cdot|$ denotes set cardinality and it is assumed that $T_\ell(O_i, O_j) = 0$. Alternatively, $d_\ell(O_i, O_j)$ is the number of object pairs in S that are more similar than $\{O_i, O_j\}$ but belong to different subsets in ℓ. The index α_ℓ is defined as

$$\alpha_\ell = \sum_{i<j} d_\ell(O_i, O_j) / \max \sum_{i<j} d_\ell(O_i, O_j) \, , \tag{3.2}$$

where the "max" is taken over all possible rank orderings of the object pairs, and both summations are taken over object pairs $\{O_i, O_j\}$ such that the indicator $T_\ell(O_i, O_j) = 0$. Although the index in (3.2) is intuitively reasonable, an alternative index is available that turns out to be a simple function of α_ℓ when the original proximity values contain no ties. If $r(O_i, O_j)$ is the rank assigned to the object pair $\{O_i, O_j\}$ on the basis of proximity, where similar object pairs are assigned the lower ranks, then the γ measure of rank correlation between $r(O_i, O_j)$ and $T_\ell(O_i, O_j)$ developed by Goodman and Kruskal [7] is equivalent to $1 - 2\alpha_\ell$. The γ index varies from -1.00 to $+1.00$, with 1.00 being obtained if and only if the partition ℓ is "perfect" in the sense that no object pair is contained in one subset of ℓ that is less similar than an object pair spanning two subsets. In fact, among all the standard rank correlation indices, this property appears unique to γ and partially justifies its use in the clustering context as a way of measuring the adequacy of any partition in a hierarchy.

The γ measure also inherits the standard probablistic interpretation with respect to a specific data set that was originally advanced as justification for the index (see [7] for a more extensive discussion). Thus, the γ measure is used here as the statistic G for which null and nonnull distributions are obtained.

Following Ling [12], a reference distribution for the γ measure at a particular partition level can be obtained for each monotone invariant clustering strategy by specifying a hypothesis of randomness in the assignment of the $n(n-1)/2$ proximity ranks for the object pairs. More precisely, all $[n(n-1)/2]!$ assignments of proximity ranks are assumed equally likely and the partition ℓ_m, say, is obtained for each such assignment by some specific clustering strategy; the γ values are then tabulated to form the sampling distribution for γ. Although the sampling distribution for the γ values at a given partition level m can be approached through simple randomization, it is just as important to develop some concept of a nonnull sampling distribution for γ when the proximity values have not been assigned completely at random. A clear characterization of a nonnull distribution for γ would provide a way of evaluating the various clustering procedures with respect to their ability to detect nonrandomness, i.e., power against an alternative, through the γ statistic. In order to develop this idea in detail, it is necessary to define the possible "true" structures, i.e., alternative hypotheses, that could exist at a given partition level.

The number of possible alternative hypotheses at the mth partition level of n objects can be determined as follows: Let n be the number of objects in the set S, and consider the partition constructed by a clustering procedure at level m. Since level m contains $n - m$ subsets, there are $R(n, n-m)$ different partitions that could be obtained where $R(n, n-m)$ is the Stirling number of the second kind, e.g., when $n = 12$ and $m = 9$, $R(12, 3) = 86{,}526$. The breakdown of this total over the specific partition types is listed in Table 1, where it should be observed that the possible types of partitions are not equally likely.

1. Partition Types at Level 9 for n = 12

Partition type	*No. of realizations*
{10,1,1}	66
{9,2,1}	660
{8,3,1}	1980
{8,2,2}	1485
{7,4,1}	3960
{7,3,2}	7920
{6,5,1}	5544
{6,4,2}	13860
{6,3,3}	9240
{5,5,2}	8316
{5,4,3}	27720
{4,4,4}	5775
Total	86526

For purposes of investigating the power of our two hierarchical clustering methods the following paradigm was used to define the matrix of object-pair proximity measures that explicitly characterize the alternative hypothesis of interest.

(a) A "true" structure is defined by selecting one specific type of partition at level m_0, say ℓ_{m_0}. The partition itself can be characterized by the dichotomous ranking

$$T(O_i, O_j) = \begin{cases} 0 \text{ if } O_i \text{ and } O_j \text{ belong to the same subset} \\ \text{in } \ell_{m_0}; \\ 1 \text{ otherwise.} \end{cases}$$

(b) "Noise" of a normal form is imposed on $T(\cdot, \cdot)$ with mean zero and variance σ^2 providing a symmetric proximity function $H(\cdot, \cdot)$ on $S \times S$:

$$H(O_i, O_j) = T(O_i, O_j) + \epsilon ,$$

where $\epsilon \sim N[O, (\sigma_\epsilon)^2]$, $i < j$.

The imposition of error on the dichotomous ranking can be interpreted as an attempt to "hide" the "true" structure with the degree of distortion measured by the magnitude of $(\sigma_\epsilon)^2$, an approach used previously by Baker [2].

(c) Since any clustering method constructs a partition at each level m, a dichotomous ranking $C_m(\cdot, \cdot)$ of the object pairs at level m is constructed as in (a). A measure of the correspondence between a constructed partition and the proximity measures $H(\cdot, \cdot)$ produced by adding noise to a "true" structure will be defined by a γ statistic between the ranking $C_m(\cdot, \cdot)$ and the rank ordering of object pairs obtained from $s(\cdot, \cdot)$; this particular index will be denoted by γ_{HC_m}.

Using each clustering strategy, the nonnull distribution of the γ statistic at any level m for a specific type of true partition imbedded at level m_0 can be obtained by repeating Steps (b) and (c) for a large number of random samples and calculating γ_{HC_m} for each. Power is the proportion of the distribution of the γ_{HC_m}'s that exceeds the α cut-off point of the null distribution or, within our context, the probability of rejecting a hypothesis of randomness when the true alternative hypothesis has been obscured by a level of "noise" defined by $(\sigma_\epsilon)^2$. Consequently, as $(\sigma_\epsilon)^2$ approaches ∞, the distribution of γ at any level m should approach the sampling distribution for γ at that partition level under the hypothesis of randomness, and power should approach the chosen significance level. Conversely, as $(\sigma_\epsilon)^2$ gets near to 0, the power should get very close to 1.00, at least for that level at which the true structure was defined.

Given the literature on the complete-link and single-link clustering procedures, it is reasonable to conjecture that the single-link strategy is sensitive to "true" partitions containing large object classes, and the complete-link strategy is sensitive to "true" partitions containing object classes of a more uniform size (see [11]). Thus, if these intuitions are correct, clustering procedures will be differentially sensitive to alternatives of a particular form specified by the sizes of the partition classes; or, alternatively, the power of our two clustering procedures should be a function of the type of "true" structure considered.

4. PROCEDURES AND RESULTS

The approach discussed here for investigating the power of the single-link and complete-link hierarchical clustering procedures was implemented as follows. All partitions of twelve objects at Level 9 given in Table 1 were used to construct object pair matrices containing 1's and 0's which thus represent the "true" structures. The sampling distribution of γ under the null hypothesis of randomness was constructed for Level 9 by adding error with mean 0 and arbitrarily large variance $(\sigma_\epsilon)^2$ to each element in an object pair matrix consisting of all 0's. The resulting matrix of proximity measures was then analyzed by both the single-link and complete-link hierarchical clustering method and, in both instances, the value of γ calculated for the ninth partition level. The process was then repeated 1,000 times giving the empirical cumulative null distributions of γ presented in Table 2.

It is interesting to observe that the single-link criterion yields a lower average value of γ than did the complete-link procedure; also, the variance is larger for the single-link strategy. Consequently, in a random assignment of

2. Approximate Null Sampling Distributions for γ at Level 9 Using Single-Link and Complete-Link Procedures; n = 12

Cum. proportion	*Single-link*[a]	*Complete-link*[b]
.10	.10	.33
.20	.15	.37
.30	.20	.40
.40	.24	.43
.50	.27	.46
.60	.31	.49
.70	.34	.52
.80	.38	.55
.90	.44	.60
.95	.49	.63
1.00	.70	.79

[a] Mean = .268; std. dev. = .138.

[b] Mean = .460; std. dev. = .104.

proximity values, the complete-link procedure tends on the average to produce a higher agreement as measured by γ at Level 9 between the constructed partition and the object pair proximity measures.

Approximate significance levels of .05, .10 and .20 were used in a decision rule for rejecting the null hypothesis of randomness. From Table 2, the corresponding α cut-off values of γ are .49, .44 and .38 for the single-link method and .63, .60, and .55 for the complete-link method. The empirical results were generated under the following conditions:

(a) The "true" alternative hypothesis was one of the 12 partitions listed in Table 1.

(b) The perturbed proximity measures $H(\cdot, \cdot)$ were obtained for error levels of σ_ϵ = .3, .5, .7, .9 and 1.1.

(c) At each of the possible error levels, 200 replications were made, analyzed by the complete-link and single-link procedures and a value of γ_{HC_m} obtained for each construction.

Table 3 provides the major empirical results and includes for each true partition the probability of rejecting the randomness hypothesis as a function of σ_ϵ for the two clustering procedures and the three chosen significance levels. For economy of space, data on only seven of the twelve possible partitions are given. One representative partition is listed from the subset of partitions that has the same number of objects within the largest object class; results for other members were similar.[1]

The data in Table 3 support our expectations since the complete-link technique is more powerful than the single-link for a "true" partition containing less than seven of the twelve objects within its largest class. Conversely, the single-link technique is more powerful for partitions defined by a largest object class containing more than seven of the twelve objects. Partitions containing a class of exactly seven objects produce ambiguous results and appear to be transitional alternatives. Thus, the information provided by the power tables supports the oft-stated contention that the single-link criterion is to be preferred for only a rather limited number of data structures (see [11]).

The obscuring effect of the error appended to the alternative hypothesis also clearly agrees with expectations. In all cases, increasing levels of σ_ϵ were accompanied by lower power against the specific alternative hypothesis. The differential impact of σ_ϵ on the power as a function of the clustering procedure again depended on the number of objects in the largest subset of the partition. For instance, when the number was greater than seven, the decrease in power was somewhat less for the single-link procedure; and when the number was less than seven, the decrease in power was less for the complete-link procedure. At the lowest level of σ_ϵ, both of the two clustering techniques yielded power values near unity for all partitions reported in Table 3.

3. Estimated Probability of Rejecting the Randomness Hypothesis Using Approximate Sampling Distribution of γ at Level 9; n = 12

Partition	Significance level	σ_ϵ	Clustering criterion	
			Complete-link	Single-link
{10,1,1}	.05	.3	.92	1.00
		.5	.52	.94
		.7	.19	.74
		.9	.11	.56
		1.1	.07	.29
	.10	.3	.94	1.00
		.5	.59	.95
		.7	.27	.79
		.9	.21	.69
		1.1	.11	.40
	.20	.3	.97	1.00
		.5	.73	.97
		.7	.45	.87
		.9	.38	.75
		1.1	.26	.55
{9,2,1}	.05	.3	.97	1.00
		.5	.73	.99
		.7	.31	.78
		.9	.17	.48
		1.1	.12	.24
	.10	.3	.99	1.00
		.5	.79	.99
		.7	.43	.84
		.9	.24	.59
		1.1	.18	.37
	.20	.3	1.00	1.00
		.5	.87	.99
		.7	.59	.89
		.9	.37	.71
		1.1	.27	.52
{8,2,2}	.05	.3	.99	1.00
		.5	.83	.90
		.7	.43	.62
		.9	.23	.39
		1.1	.10	.16
	.10	.3	.99	1.00
		.5	.85	.94
		.7	.53	.74
		.9	.29	.52
		1.1	.14	.31
	.20	.3	.99	1.00
		.5	.90	.92
		.7	.57	.85
		.9	.40	.68
		1.1	.30	.37
{7,3,2}	.05	.3	1.00	.97
		.5	.78	.67
		.7	.39	.39
		.9	.25	.21
		1.1	.12	.16
	.10	.3	1.00	.98
		.5	.83	.77
		.7	.48	.48
		.9	.32	.30
		1.1	.19	.27
	.20	.3	1.00	1.00
		.5	.88	.89
		.7	.62	.63
		.9	.49	.45
		1.1	.35	.41

[1] Although we actually use these estimated values as if they were exact, the sample size of 1,000 provides a 99 percent confidence interval for the 90th percentile that ranges from .56 to .61 for the complete-link value and from .39 to .46 for the single-link. Thus, the estimation of these two cut-off points is fairly precise. See [5] for the necessary formulas.

Table 3. (Continued)

Partition	Significance level	σ_ϵ	Clustering criterion	
			Complete-link	Single-link
{6,4,2}	.05	.3	1.00	.94
		.5	.86	.65
		.7	.48	.33
		.9	.26	.19
		1.1	.14	.13
	.10	.3	1.00	.96
		.5	.88	.72
		.7	.55	.42
		.9	.31	.26
		1.1	.24	.20
	.20	.3	1.00	.98
		.5	.90	.81
		.7	.67	.54
		.9	.51	.39
		1.1	.37	.30
{5,4,3}	.05	.3	1.00	.92
		.5	.87	.54
		.7	.48	.33
		.9	.31	.17
		1.1	.20	.19
	.10	.3	1.00	.94
		.5	.92	.63
		.7	.53	.42
		.9	.39	.23
		1.1	.27	.26
	.20	.3	1.00	.96
		.5	.94	.75
		.7	.69	.54
		.9	.53	.35
		1.1	.37	.36
{4,4,4}	.05	.3	.97	.96
		.5	.70	.53
		.7	.40	.22
		.9	.19	.09
		1.1	.14	.08
	.10	.3	.97	.98
		.5	.76	.65
		.7	.48	.33
		.9	.22	.15
		1.1	.20	.12
	.20	.3	.99	.99
		.5	.86	.75
		.7	.63	.46
		.9	.36	.30
		1.1	.32	.20

Table 4A presents the means and variances for the nonnull distributions of γ_{HC_9} under the same values of σ_ϵ. In terms of the mean values of γ_{HC_9} given in Table 4A, it appears that the single-link procedure will produce the higher values for only the lower error levels, $\sigma_\epsilon = .3$, .5 and .7, used for the {10, 1, 1} and {9, 2, 1} partitions. In all other instances, the complete-link strategy generates the highest average γ_{HC_9} and smaller standard deviations of γ_{HC_9} as well. It is interesting to note that the effect of increasing levels of σ_ϵ on the standard deviation of γ_{HC_9} is counter to intuition. In particular, the standard deviations of γ_{HC_9} generally increase slightly and then decrease for both the complete-link and single-link procedures as σ_ϵ increases.

5. ESTIMATION

Even though the problem of obtaining sampling distributions for the γ values at a particular partition level m can be approached through the procedures of this article, there is another important statistical problem that also needs attention once some reasonable concept of a sampling distribution and power is available. More specifically, it is important to define criteria of estimation that would be of use in evaluating whether a partition constructed by a clustering scheme is close to the "true" structure. Such criteria would allow assessment of a clustering strategy's ability to elicit whatever structure underlies the data.

As a measure of how close the partitions produced by a clustering method correspond to the "true" partitions, a γ statistic, denoted by γ_{TC_m}, can be used between the dichotomous rankings $T(\cdot,\cdot)$ and a dichotomous ranking

4. Sample Means and Standard Deviations for the γ_{TC} Statistic and the Nonnull Distribution of the γ_{HC} Statistic at Level 9; $n = 12$

Partition	σ_ϵ	Clustering criterion			
		Complete-link		Single-link	
		Mean	Std. dev.	Mean	Std. dev.
		A. γ_{HC}			
{10,1,1}	.3	.919	.140	.981	.020
	.5	.662	.162	.821	.129
	.7	.536	.120	.613	.177
	.9	.509	.106	.509	.177
	1.1	.487	.097	.405	.142
{9,2,1}	.3	.959	.088	.963	.061
	.5	.727	.146	.774	.110
	.7	.578	.124	.580	.148
	.9	.509	.118	.469	.156
	1.1	.494	.107	.382	.148
{8,2,2}	.3	.979	.048	.922	.138
	.5	.783	.139	.709	.160
	.7	.605	.147	.531	.155
	.9	.528	.125	.440	.151
	1.1	.486	.113	.363	.134
{7,3,2}	.3	.982	.018	.927	.142
	.5	.769	.159	.611	.205
	.7	.595	.134	.445	.176
	.9	.533	.136	.376	.156
	1.1	.501	.113	.342	.151
{6,4,2}	.3	.979	.027	.921	.161
	.5	.783	.137	.603	.228
	.7	.616	.145	.414	.197
	.9	.546	.132	.354	.163
	1.1	.507	.120	.317	.147
{5,4,3}	.3	.981	.026	.914	.179
	.5	.799	.130	.576	.248
	.7	.619	.140	.409	.196
	.9	.556	.126	.322	.166
	1.1	.522	.121	.322	.180
{4,4,4}	.3	.886	.109	.884	.140
	.5	.716	.148	.519	.215
	.7	.595	.139	.350	.174
	.9	.506	.124	.290	.152
	1.1	.498	.118	.282	.140

Table 4. (Continued)

Partition	σ_ϵ	Clustering criterion			
		Complete-link		Single-link	
		Mean	Std. dev.	Mean	Std. dev.
		B. γ_{TC}			
{10,1,1}	.3	.980	.054	1.000	.000
	.5	.849	.201	.972	.127
	.7	.627	.342	.871	.239
	.9	.500	.373	.785	.337
	1.1	.371	.370	.648	.394
{9,2,1}	.3	.995	.024	.998	.007
	.5	.890	.158	.961	.124
	.7	.684	.301	.872	.236
	.9	.535	.368	.740	.374
	1.1	.407	.377	.657	.343
{8,2,2}	.3	.997	.029	.990	.020
	.5	.918	.153	.944	.102
	.7	.726	.281	.844	.263
	.9	.530	.351	.735	.331
	1.1	.463	.363	.636	.402
{7,3,2}	.3	1.000	.001	.989	.028
	.5	.895	.205	.881	.201
	.7	.724	.283	.726	.328
	.9	.588	.350	.599	.373
	1.1	.427	.344	.517	.396
{6,4,2}	.3	.998	.010	.991	.030
	.5	.941	.118	.863	.200
	.7	.788	.256	.612	.349
	.9	.606	.334	.462	.370
	1.1	.541	.333	.322	.388
{5,4,3}	.3	1.000	.004	.984	.045
	.5	.943	.137	.771	.280
	.7	.763	.242	.534	.349
	.9	.582	.335	.369	.354
	1.1	.495	.345	.261	.324
{4,4,4}	.3	.981	.045	.985	.042
	.5	.894	.154	.788	.253
	.7	.752	.258	.486	.352
	.9	.564	.315	.282	.333
	1.1	.441	.348	.174	.290

$C_m(\cdot,\cdot)$ of the object pairs at level m constructed by a clustering procedure. Table 4B presents additional comparisons using the means and variances of γ_{TC_9} for the single-link and complete-link procedures and for the selected "true" partitions and levels of σ_ϵ from the example of Section 4. For all "true" structures based on twelve objects containing largest subsets of fewer than seven members, the complete-link procedure generally produces a partition that, on the average, is closer to the original dichotomous ranking. The advantage of the complete-link over the single-link procedure increases as σ_ϵ increases, and at the partition {4, 4, 4} a mean of .441 was obtained for the former and .174 for the latter. Thus, besides being more powerful for "true" partitions containing less than seven objects in the largest class, the complete-link strategy estimates the "true" structure better than the single-link for these partitions as well. Conversely, for the "true" structures that are detected best by the single-link procedure, the single-link estimates the true structure better, i.e., for those alternatives having more than seven of the twelve objects in the largest subset of the partition.

The mean and standard deviations of γ_{TC} depend on the value of σ_ϵ in such a way that as σ_ϵ increases, $\bar{\gamma}_{TC}$ decreases and $\sigma_{\gamma_{TC}}$ increases. When there are fewer than seven objects in the largest object class of a partition, the complete-link results tend to be less variable and, alternatively, the single-link results tend to be less variable when a largest class contains seven or more objects.

6. SUMMARY

The empirical data suggest that the complete-link and the single-link clustering procedures are differentially sensitive to particular partitions of objects imbedded in the proximity values. These results indicate that the single-link procedure is likely to reject the randomness hypothesis and estimate the "true" partition better when the "true" partition includes a single large subset. When there were fewer than seven objects in the largest subset, however, the complete-link procedure yields higher power and better estimates. The reader should be reminded, however, that the study was based only on the sample size of twelve and consequently, the power of the two clustering procedures is unknown for other sample sizes. In particular, the discussion here using seven objects in the largest subset as a "breakpoint" is a function of the sample size of twelve that is considered.

These conclusions are not surprising but do add a firm empirical base to previously reported empirical observations (see [11]). Although this article studies power only at Level 9 of the hierarchy, extension of these procedures to other levels of the hierarchy is a simple though somewhat expensive and tedious procedure. If such a program were carried out for a range of object set sizes, however, fairly complete power and estimation tables could be developed for the complete-link and single-link clustering strategies. This information would allow a researcher who had some idea of the type of partition of interest in a given hierarchy and the amount of error in the data to choose the most appropriate clustering procedures.

From a theory point of view, the use of partitions of objects at a specific level of the hierarchy appears promising since it reduces the power and estimation problem to a more tractable single level of a hierarchy. Although the combinatorial problems are considerable, conceivably it would be possible to work towards a concept of power for complete hierarchies by combining the results obtained for individual levels.

[*Received May 1974. Revised October 1974.*]

REFERENCES

[1] Anderberg, M.R., *Cluster Analysis for Applications*, New York: Academic Press, 1973.

[2] Baker, F.B., "Stability of Two Hierarchical Grouping Techniques; Case I: Sensitivity to Data Errors," *Journal of the American Statistical Association*, 69 (June 1974), 440–5.

[3] Bradley, J.V., *Distribution-Free Statistical Tests*, Englewood Cliffs, N.J.: Prentice-Hall, 1968.

[4] Collier, R.O. and Baker, F.B. "Some Monte Carlo Results on Power of the *F*-Test Under Permutation in the Simple Randomized Block Design," *Biometrika*, 53, Nos. 1 and 2 (1966), 199–203.

[5] Conover, W.J., *Practical Nonparametric Statistics*, New York: John Wiley and Sons, Inc., 1971.

[6] Fisher, R.A., *Design of Experiments*, Edinburgh: Oliver & Boyd, Ltd., 1949.

[7] Goodman, L.A. and Kruskal, W.H., "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49 (December 1954), 732–64.

[8] Hubert, L., "Some Extensions of Johnson's Hierarchical Clustering Algorithms," *Psychometrika*, 37 (September 1972), 261–74.

[9] ———, "Approximate Evaluation Techniques for the Single-link and Complete-link Hierarchical Clustering Procedures," *Journal of the American Statistical Association*, 69 (September 1974), 698–704.

[10] Kempthorne, O., *The Design and Analysis of Experiments*, New York: John Wiley and Sons, Inc., 1952.

[11] Lance, G.N. and Williams, W.T., "A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems," *The Computer Journal*, 10 (February 1967), 373–80.

[12] Ling, R.F., "A Probability Theory of Cluster Analysis," *Journal of the American Statistical Association*, 68 (March 1973), 159–69.

[13] Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy*, San Francisco: W.H. Freeman and Co., 1973.

[14] Rao, C.R., *Advanced Statistical Methods in Biometric Research*, New York: John Wiley and Sons, Inc., 1952.