Pattern Recognition Vol. 11, pp. 235–254.
Pergamon Press. Ltd. 1979. Printed in Great Britain.
© Pattern Recognition Society.

0031-3203/79/0801-235 $02.00/0

VALIDITY STUDIES IN CLUSTERING METHODOLOGIES*

RICHARD DUBES and ANIL K. JAIN
Department of Computer Science, Michigan State University, East Lansing, MI 48824, U.S.A.

(*Received* 19 *October* 1978; *in revised form* 16 *February* 1979; *received for publication* 13 *March* 1979)

Abstract – Clustering algorithms tend to generate clusters even when applied to random data. This paper provides a semi-tutorial review of the state-of-the-art in cluster validity, or the verification of results from clustering algorithms. The paper covers ways of measuring clustering tendency, the fit of hierarchical and partitional structures and indices of compactness and isolation for individual clusters. Included are structural criteria for validating clusters and the factors involved in choosing criteria, according to which the literature of cluster validity is classified. An application to speaker identification demonstrates several indices. The development of new clustering techniques and the wide availability of clustering programs necessitates vigorous research in cluster validity.

Clustering   Cluster validity   Hierarchical structure   Clustering tendency   Compactness   Isolation
Intrinsic dimensionality   Global fit

1. INTRODUCTION

Clustering algorithms are valuable tools in exploratory data analysis and pattern recognition studies since they help one "look" at the data and ascertain characteristics of its structure by organizing data into subgroups, or clusters.† If one has a great deal of experience with a particular clustering method and some prior information about the data being clustered, the results of a clustering algorithm can confirm or deny assertions about the data and suggest subsequent analyses. When such qualitative and *ad hoc* information is sufficient for the user's needs, interpreting the results of clustering algorithms becomes a personal matter in which intuition and insight are dominant. However, the user of a clustering algorithm is often unsure about the data and has little experience with a particular type of data or a particular clustering method. Lack of information about the data is often the motivation for clustering the data in the first place. In this case, the user searches for objective meaning and needs quantitative measures of significance for evaluating clustering structures.

This paper summarizes in a semi-tutorial manner procedures available in the literature for quantitatively evaluating the results of clustering methods without regard for the subject matter of the data. We carefully enumerate the factors involved in establishing the concepts of "valid" clustering structures, clusterings and individual clusters and categorize the ways in which objective meaning can be assigned to the results of clustering algorithms. The term "cluster validity" will refer to this entire range of activities. Our objectives are to make users of clustering algorithms aware of existing techniques for validating clusters and cognizant of the limitations on existing knowledge. The broad spectrum of journals containing information relevant to cluster validity make these techniques and limitations very difficult to track down. In fact, our brief list of references contains over 30 different journals. We also hope to motivate researchers in pattern recognition to attack problems in cluster validity.

The purview of our study can be demonstrated on data from a speaker recognition study consisting of 25 samples (or patterns) of choral speech.‡ Dissimilarities between all pairs of patterns in this data set were computed with the Manhattan distance metric in a 72-dimensional pattern space, the 72 features being log magnitudes of energy spectra. The dendrograms representing the results of the single-link and complete-link clustering methods[34] are shown in Fig. 1. Several questions can be posed. Are the clustering structures displayed in Fig. 1 real or merely artifacts of the algorithms? Are any of the partitionings obtained by cutting the dendrograms appropriate summaries of the data? Which, if any, of the individual clusters are valid? Which of the two hierarchical structures is better? Do others exist which are still better? A long

* Research supported by National Science Foundation Grant ENG 76-11936A01.

† For background information on clustering methodology see.[2, 14, 15, 18, 25, 28, 31, 39, 73] A preliminary version of this paper was presented earlier.[19]

‡ Choral speech[12, 79] is obtained by dividing a segment of speech from a single speaker into contiguous segments and averaging the segments.

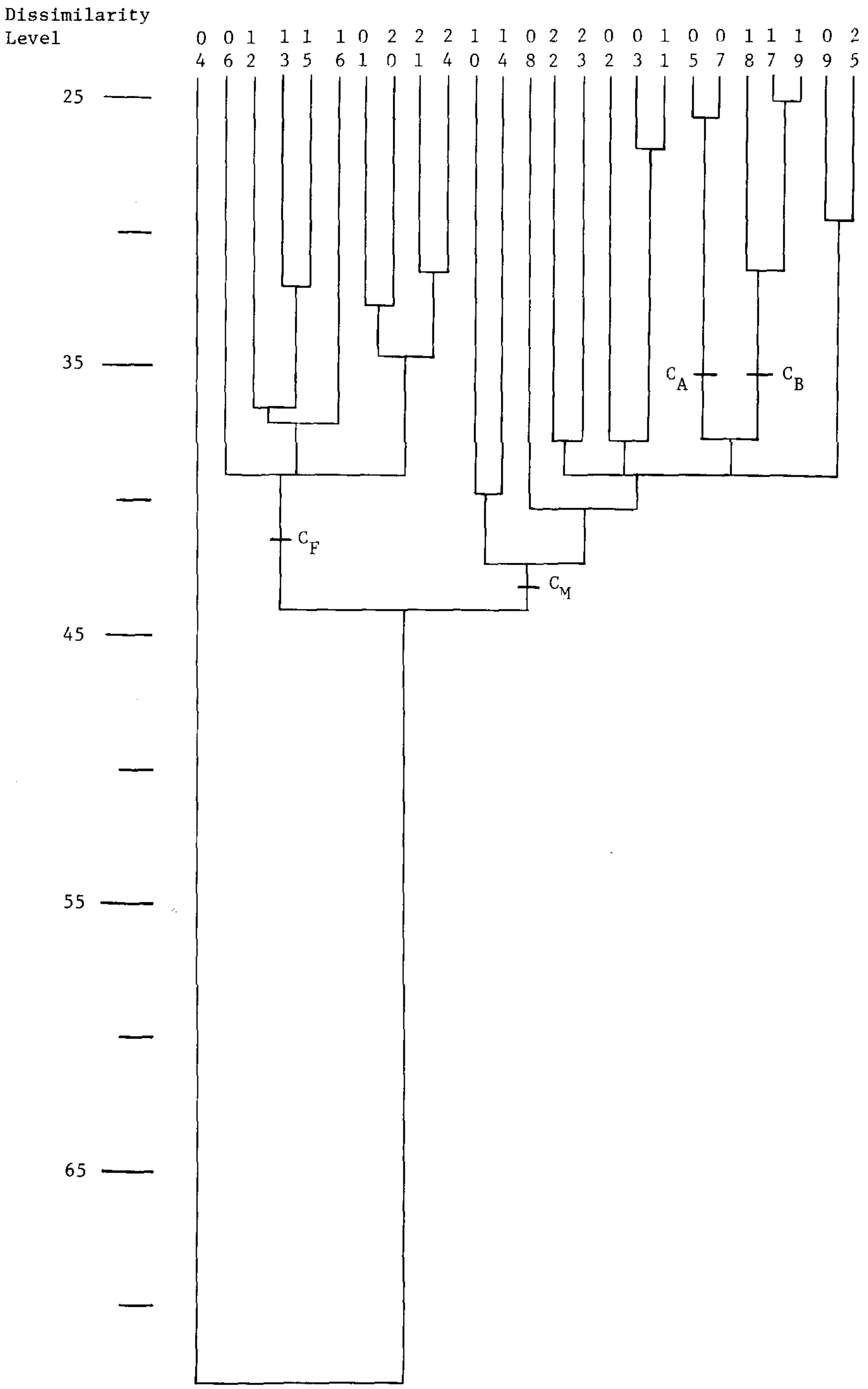

Fig. 1(a). Single link dendrogram (distance).

list of such questions can be posed. We demonstrate various measures of cluster validity for this data set.

One way to qualitatively evaluate the hierarchies is to impose *a priori* categories on the data. For example, the nine patterns in cluster C_F indicated in Figs. 1(a) and (b) all represent female speakers. The 14 patterns in cluster C_M represent male speakers. Pattern 4 is a female speaker. Are the clusters by sex valid? The experiment involved 11 speakers. Clusters C_A and C_B contain all the samples from two specific speakers. Is the formation of these clusters significant or merely a chance occurrence? Finally, the samples are in three languages; all the speakers are multilingual. For example, patterns 2, 6, 9, 12, 16 and 19 are all recorded in Italian. Is the absence of an "Italian" cluster good evidence that the data are language independent?

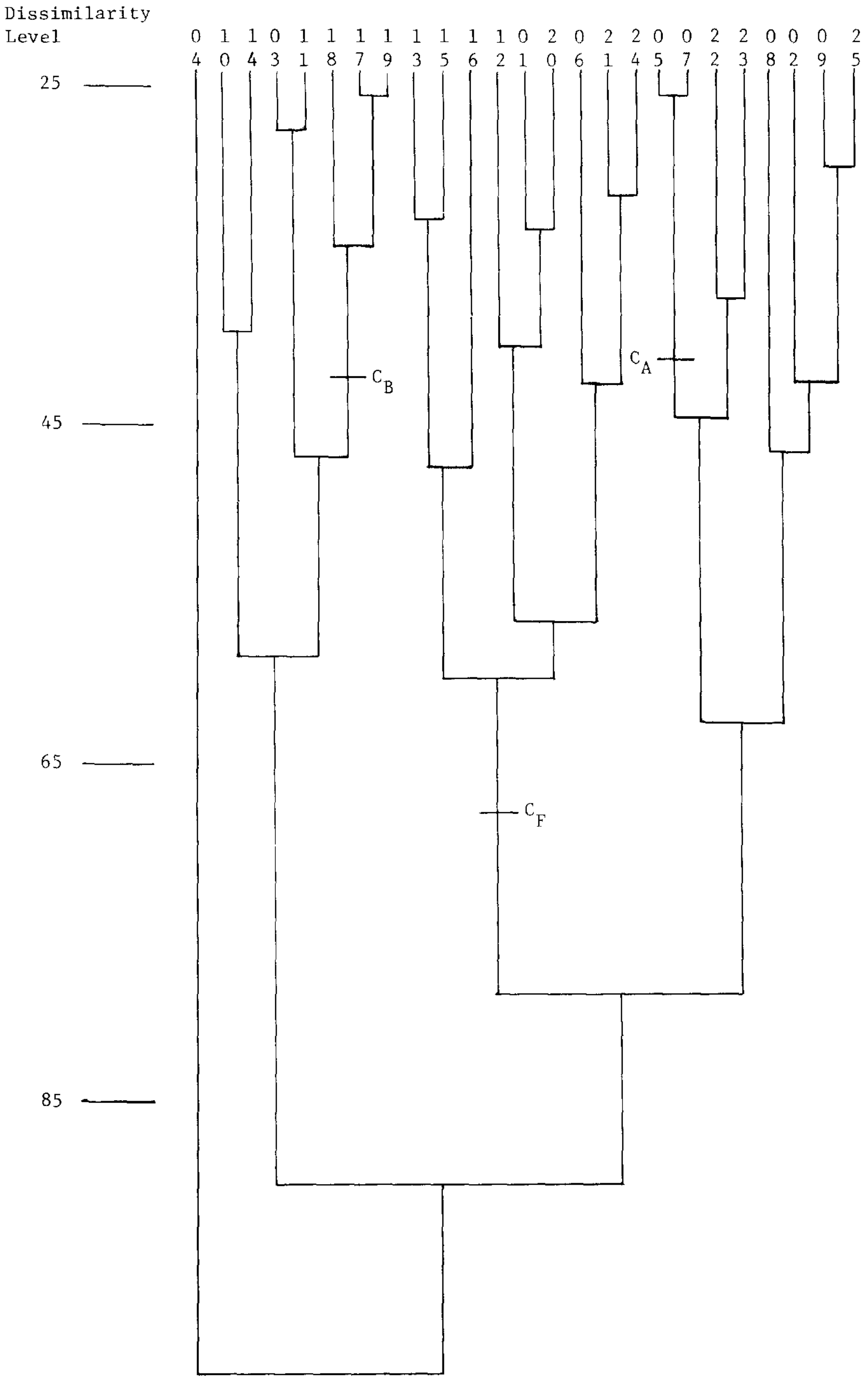

Fig. 1(b). Complete link dendrogram (distance).

These are the types of questions indices of cluster validity should answer.

As a second example, a study of computer resource usage was conducted on a general purpose digital computer. A total of 300 patterns were recorded, each pattern characterizing a job step with 11 features such as CPU time, disk I/O time, etc. for a particular time slice. Following a similar study by Agrawala,[1] a mean-square clustering algorithm based on the Friedman–Rubin[27] algorithm called CLUSTER[18] was used to investigate the data structure. The output of such a clustering algorithm consists of statistical tables such as cluster centers, cluster populations, distances between cluster centers, average distances within cluster centers and the like. No pictorial output is available for visually assessing the relative merits of

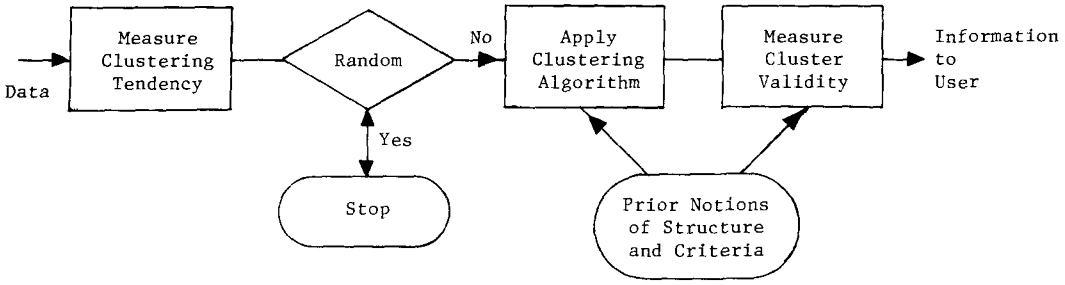
Data
Measure Clustering Tendency
Random
No
Yes
Stop
Apply Clustering Algorithm
Measure Cluster Validity
Information to User
Prior Notions of Structure and Criteria

Fig. 2. Scope of cluster validity problem.

the clusters. For reasons discussed later, the many statistical tests of significance developed in multivariate analysis of variance do not provide quantitative measures of cluster validity. Clustering algorithms will generate a clustering whether one really exists or not. Many of the questions posed above can be repeated for this example. The task of Cluster Validity is to separate the artifacts from the structure.

The fundamental questions we address can be stated from a user's point of view as: are the clusters and structures generated by a clustering method or algorithm significant enough to provide evidence for hypotheses about the phenomenon being studied? It is interesting to note how the literature in the four areas most concerned with cluster analysis treat it.

The engineering literature[1,29,32,40,43,44,55,69,78,88] largely ignores cluster validity, being content to display new algorithms and to justify the results by demonstrating that the algorithm "works" on a few mathematical structures specified *a priori*, often in two dimensions. The literature in the behavioral sciences[8,10,26,34–36,38,53,59,66,75,77,84] shows concern about cluster validity but is willing to accept a clustering if it "makes sense" in an application, attributing severe inconsistencies to "noise". The applied statistics literature[5–7,15,21,27,33,42,46–48,50,51,67,82] imbeds cluster validity questions in a hypothesis testing framework, which is a step in the right direction. Unfortunately, the type of randomness assumptions required to work out sampling distributions are seldom appropriate in actual applications. Literature in the biological and natural sciences,[14,17,24,39,54,56,61,62,64,68,70,72–74,76,77,83] especially numerical taxonomy, has been most attentive to questions of cluster validity. The importance of these questions and the difficulty of the problem have been pointed out several times.[89,90]

Potential solutions to the problem of validating clusters have been hampered by disagreement on what "cluster" means and on the meaning of validity, by inappropriate statistical models chosen only for the availability of statistical tables and by the difficulty of the underlying problem. Applied scientists have either accepted *ad hoc* methods with little theoretical justification, or borrowed statistical techniques with little regard to their robustness. One can justify such attitudes because cluster analysis is supposed to be a "tool for discovery", a means for suggesting future studies, rather than an end in itself.[2] Our position is that the applications have outrun the theory and that much more attention must be paid to the fundamental question of cluster validity if cluster analysis is to make a substantial contribution to data analysis. We believe that clustering methodology ought to be available to a broad range of engineering and scientific studies. Since clustering studies are exploratory in nature, the user must be able to check his data for clustering tendency, select the algorithm appropriate to his data and research objectives, evaluate the results, establish their validity and form hypotheses about his data as rapidly and as easily as possible.[18]

2. PROBLEM DEFINITION

The first step in our critical review of cluster validity is to stratify the problem in two ways – by data and by imposed structure. We view the problem as in Fig. 2. Data are first checked for clustering tendency. Only if the data tend to be non-random is clustering attempted. The validation process judges the success of an algorithm in imposing a structure as well as the suitability of the structure for the data.

The specific notions of "data" and "structure" used in this paper are defined in Sections 2.1 and 2.2 to subdivide our review of cluster validity. Fixing these two parameters defines a situation and we will study specific solutions to questions of cluster validity in such situations.

2.1 *Data*

Clustering begins with a set of *n* patterns, or data items, or objects, or individuals, or cases, or operational taxonomic units. Inferences about the generation of these patterns are to be drawn from the results of a clustering algorithm. Information about the patterns is presented either as a pattern matrix or as a proximity matrix as indicated in Fig. 3. The mode of information presentation is usually determined by the subject matter.

In a pattern matrix each pattern is described by a set of *d* measurements or features. The features are classified according to data scales. Our notions of "distance" and of geometry are strongly influenced by data scales. For example, nominal scales are seen as binary-valued features and patterns are pictured as

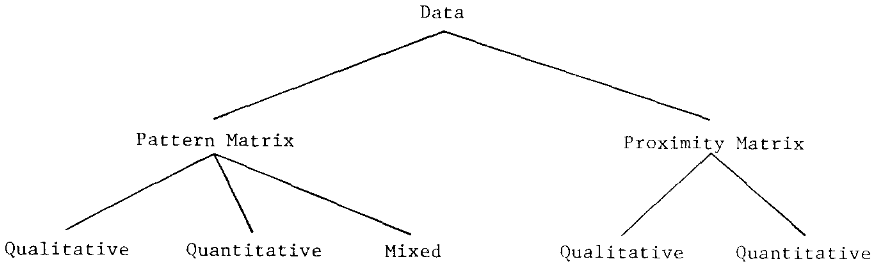

Fig. 3. Data types.

vertices of a hypercube in the d-dimensional pattern space. For quantitative data scales, Euclidean distance serves as a natural measure of separation between patterns, primarily because it fits with the geometric picture and most researchers feel comfortable with it. Minkowski metrics have also been used. When some of the features are qualitative and some are quantitative, the geometric picture loses consistency. One's mental image of a pattern matrix influences the interpretation of cluster validity measures. Here, we adopt the paradigm common to most pattern recognition studies and assume that all features are measured with infinite precision, that the patterns are all distinguishable, and that the number of patterns exceeds the number of features by a factor of five or more. The infinite precision assumption rules out situations where patterns can occur only at the intersections of grid lines and are allowed to "pile up", as studied by Kittler[(43)] and Zahl,[(87)] for example.

A proximity matrix is an $n \times n$ matrix in which each row (and column) represents a pattern. The entries are values of an index of similarity, such as correlation, or of dissimilarity, such as distance. A proximity matrix can be the basic expression of the raw data, as when collecting data from human subjects, by requiring all possible comparisons between the n patterns or can be computed from a pattern matrix. The entries can be recorded either on a quantitative scale, such as a distance in a feature space, or on a qualitative scale, such as when rank orders are used.

The utility of "seeing" the patterns as points in a space, preferably two-dimensional, has motivated the development of multidimensional scaling methods for creating a pattern matrix to characterize the given proximity matrix in the sense that distance in the hypothesized pattern space corresponds to the index of proximity observed in the proximity matrix.[(63,71)] We do not treat the validity of such representations here or consider questions of "metric determinancy".

The clustering method and measure of cluster validity used for a proximity matrix is largely determined by the data scale. An interval scale contains more detailed information than an ordinal scale and measures of cluster validity must take advantage of it.

2.2 *Imposed structure*

In this section, "structure" means the form of the results of a clustering method. The components of our concept of structure are interrelated in Fig. 4. The usual justification for the structures adopted in a clustering problem is pragmatic. The structures provide information and are generated by the clustering algorithms with which we are dealing. The two most popular types of imposed clustering structures are "partitional" and "hierarchical".

A *partitional structure* organizes the patterns into a small number of clusters by labelling each pattern in some way. We can ask about the validity or significance of a particular cluster with respect to its environment or about the goodness of the clustering, the set of clusters, as a whole. A pattern matrix is usually clustered in this way. We try to label the patterns which are "like" one another in the same way and label "unlike" patterns differently.

The fuzzy set concept, introduced by Zadeh[(86)] in 1965, has been applied to describe partitional structures.[(4,9)] In fuzzy clustering, each pattern is allowed to belong to several clusters with a measure of "belongingness" for each cluster. In conventional clustering, each pattern belongs to exactly one cluster in a partition. Several clustering criteria have been proposed for fuzzy partitions, similar to the mean-square criterion in conventional clustering. Bezdek[(9)] and Backer[(4)] have proposed significance measures for fuzzy partitions. This paper does not cover fuzzy clustering or the "flat" clusters studied by Day.[(17)] These are examples of overlapping structures[(3,39)] which are excluded from our study.

A *hierarchical structure* consists of a sequence of clusterings. We will treat the case when each clustering is a partition and the partitions are nested. Jardine and Sibson[(39)] cover more general structures. Depending on the algorithmic approach taken, a hierarchical structure begins with n clusters, one per pattern, and grows a sequence of clusterings until all n patterns are in a single cluster, or begins with one cluster containing all n patterns and successively divides clusters until n clusters are obtained. The nesting relationship between clusters is at least as important as the particular clusterings achieved.

We contend that validating the results of imposing a structure on data with a clustering method requires clear definitions of the following four structural criteria. These criteria are not independent and must be blended into a workable methodology.

(1) Compactness criterion: measures the inner

strength, or concentration or cohesion or uniqueness of an individual cluster with respect to its environment.

(2) Isolation criterion: measures the distinctiveness or separation or gaps between a cluster and its environment.

(3) Global fit criterion: measures the accuracy with which the structure describes the relationships between clusters, as well as the extent to which all the clusters are individually valid.

(4) Intrinsic dimensionality criterion: determines the "shape" of a cluster and provides information about representing the patterns in a cluster. A review of intrinsic dimensionality literature is available.[57]

There is a need to establish a methodology whereby one can incorporate specific realizations of the structural criteria into a program for exploratory data analysis. We attack this problem by fixing the data type and the type of imposed structure. The five basic factors considered below materially affect the questions posed and the answers expected. A consideration of these five factors should simplify the task of choosing criteria and clarify one's thinking about the type of questions that can be reasonably posed.

(i) *The null hypothesis.* This is a statement about the meaning of "no clustering", often expressed as a concept of randomness or the antithesis of clustering. The two null hypotheses treated in this paper will be called the "Random Graph Hypothesis" and the "Random Position Hypothesis". These are the only two null hypotheses that have been studied in the literature. The first is applicable in studies involving symmetric proximity matrices whose entries are rank orders. For example, an $n \times n$ ordinal dissimilarity matrix has zeros on the diagonal and the numbers $1, 2, \ldots, n(n-1)/2$ in the upper triangle without ties; the most similar pair of data items has rank 1. The Random Graph Hypothesis is that all $[n(n-1)/2]!$ such matrices are equally likely. Alternatively, imagine connecting all $n(n-1)/2$ pairs of nodes in an n-node graph in random order with the order of connection being the dissimilarity measure. The Random Position Hypothesis views the n patterns as independent samples from a d-dimensional distribution, such as uniform over a hypercube or hypersphere or unimodal Gaussian.

(ii) *The ideal cluster.* This factor establishes the user's prior conception of what a cluster means and sets the goal for a clustering method. One can formulate an idea of cluster for an assumed mathematical model for data generation or from prior work in the subject matter. For example, one can picture each cluster as a single point to which measurement noise has been added. A reasonable idea of a cluster would be a spherical or hyperellipsoidal swarm of patterns. The clustering method chosen will depend on the notion of an ideal cluster. For example, the single link method generates loosely connected or straggly clusters, while the complete link method generates tight clusters. McQuitty,[53] Van Rijsbergen[81] and Day[17] all use the notion of an ideal cluster to motivate work in cluster validity.

(iii) *Sample size.* Increasing the number of patterns can increase the confidence in a particular structure. It can also make the computational burden overwhelming. Criteria which are valid for, say, $n < 50$ can become computationally unreasonable for n of 5000. When the number of features in a pattern matrix is three or less, one can visually inspect the patterns and make *ad hoc* judgments about structure. If the number of features is large, one clustering strategy is to project the patterns to two dimensions and cluster by eye. This raises the questions of intrinsic dimensionality and suitable data representation. Reducing the dimensionality can have the effect of increasing sample size.

(iv) *Details of imposed structure.* Each clustering method impresses its own set of restrictions. Sometimes the restrictions are implicit in the definition of the clusters. Many clustering algorithms always find clusters which are ball-shaped. Others always place the two patterns which are closest in the same cluster. In addition to such fixed restrictions, clustering algorithms often require the user to supply limiting values of various kinds. For example, the choice of connectivity measure affects the type of hierarchical structure that can be obtained.

(v) *Ultimate use.* Clustering methods have been touted as tools for discovery, rather than ends in themselves. Thus, we might search for more descriptive structures when evaluating a clustering than we would, for example, when evaluating a statistic for estimating the mean of a binomial distribution or a decision rule for discriminating between two Gaussian distributions.

3. SURVEY OF LITERATURE ON CLUSTER VALIDITY

This section summarizes most of the literature in cluster validity that is applicable to real problems. We concentrate on specific test statistics having sampling distributions which are known or which can be approximated and emphasize the quantitative results currently available. Our interest is in validating clusters, not creating them, so many intuitively appealing indices for forming clusters are ignored. The presentation is organized according to the following four questions (see Fig. 5).

Is the data matrix random? (Section 3.1). Unless some evidence exists that the data tend to cluster, one has no basis for imposing any cluster structure on the data.

How well does a hierarchy fit a proximity matrix? (Section 3.2). A high degree of global fit between dendrogram and proximity matrix is necessary if all the clusters are to be meaningful.

Is a partition valid? (Section 3.3). The partition may be formed from a pattern-matrix description of the

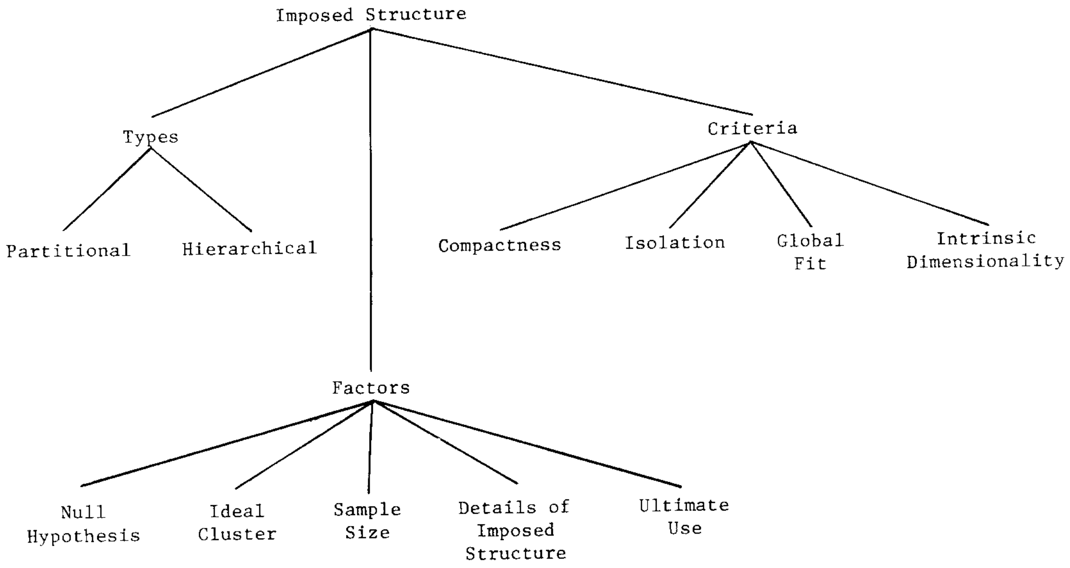

Fig. 4. Structural considerations in cluster validity.

data or obtained from one level of a hierarchy constructed from a proximity matrix.

Which individual clusters appearing in a hierarchy are valid? (Section 3.4). Here we fit a hierarchical structure, or dendrogram, to a proximity matrix and ask which, if any, of the clusters formed are "real".

Trying to answer these four questions organizes the literature on cluster validity in a reasonable way. The questions are not mutually exclusive and a methodology for incorporating answers into what one might call a program for testing cluster validity will not be apparent. The problem of using the dependence among features to identify the causal relationships which exist in the process that generates the data is related to clustering, as pointed out by Borucki *et al.*,[13] Bonner[11] and White and Lewinson.[82] Our study does not include these "causality clusters".

Our emphasis here is on external criteria that are independent of the data being studied. One strategy taken in clustering studies sidesteps the entire question of cluster validity by choosing a criterion, say, of global fit, creating a clustering method that optimizes the criterion, at least locally, and assuming that any clustering produced under such a method must be valid. We reject this strategy for two reasons. First, a clustering method always finds clusters, whether or not they are real. For example, if the patterns lie along two long parallel lines, a mean-squared algorithm will likely cut the lines rather than group the patterns on each line. Second, this strategy uses some criterion for

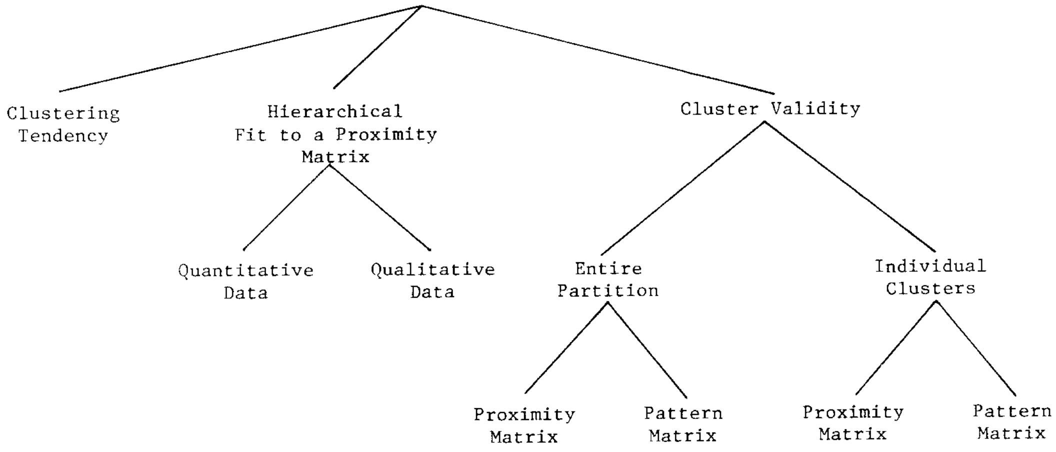

Fig. 5. Classification of validity studies.

forming the cluster and then implicitly uses the same criterion for testing the validity of the cluster while ignoring the special circumstances of its formation. Our interest is in *a priori* criteria that are established without reference to the data.

3.1 *Measure of clustering tendency*

This section reviews four approaches to the problem of determining whether the data are random. The fact that the data are not random does not mean that a clustering structure is appropriate, but it is certainly foolish to impose a clustering structure on data known to be random.

The first three approaches require a rank-order dissimilarity matrix $[r_{ij}]$ and are based on the concept of a random graph.[34] We begin with a set of n nodes, one per pattern. A threshold graph $G_{n.v}$ is an undirected graph containing n nodes and v edges with edge (i,j) being entered if $r_{ij} \leq v$. Under the Random Graph Hypothesis, these edges are entered randomly.

(a) Fillenbaum and Rapoport[26] and Rapoport and Fillenbaum[59] suggest three tests for examining clustering tendency in ordinal dissimilarity matrices.

The first test involves the number of edges, V, needed to connect a random graph. Knowing the distribution of this number of edges permits one to judge how many edges must be observed before deciding that the data are random. Fillenbaum and Rapoport used an asymptotic equation for the probability that $G_{n.v}$ is connected that was originally derived by Erdös and Réyni.[22] Schultz and Hubert[66] showed that the asymptotic probability was not accurate for small sample sizes. Ling[46] and Ling and Killough[50] settled the question by producing an exact equation for the probability in question, based on results of Riddel and Uhlenbeck.[60] Ling[46] adopts the Random Graph Hypothesis which requires that all $[n(n-1)/2]!$ ordinal dissimilarity matrices be equally likely. The distribution function for V is denoted:

$$P_{n.v} = \text{Prob}(V \leq v|n).$$

If v^* is observed in a particular situation as the level at which the graph for the data being studied first becomes connected, or the level at which all data items are absorbed into the same single-link cluster, then clustering tendency is tested as follows.

If $P_{n.v*} > 0.99$, evidence exists for the conclusion that the dissimilarity matrix was *not* chosen at random. The threshold 0.99 is arbitrary. The intuitive idea behind this test is that the within-cluster edges will tend to occur before the between-cluster edges when the data are clustered, thus delaying the formation of a connected graph.

As an example, the 25-sample choral speech dissimilarity matrix discussed in Section 1 was translated to a rank matrix and clustered by the single-link method. The dendrogram is given in Fig. 6. Since the single-link method uses only the rank orders of the dissimilarities to form clusters, the dendrograms of Figs. 1(a) and 6 are identical except for scale, even though they appear different. We observe $v^* = 247$. From Table 3 of Ling and Killough,[50]

$$P_{25.81} > 0.99$$

thus,

$$P_{25.247} \gg 0.99.$$

Figure 6 shows that sample 4 dominates the value of V. If sample 4 were not in the data set, the resulting 24-node graph would connect at rank 54. Since

$$0.85 < P_{24.54} < 0.9$$

the 24-node graph provides only weak evidence for the existence of clusters according to this statistic.

The second test suggested by Rapoport and Fillenbaum[59] observes the distribution of node degrees in an experimentally derived threshold graph $G_{n.v}$, n and v fixed. The degree of a node is the number of edges incident to the node. If R represents the number of edges of $G_{n.v}$ incident to a particular node, the probability mass function of R under the Random Graph Hypothesis is:

$$\text{Prob}(R=r) = \frac{\binom{n-1}{r}\binom{\binom{n-1}{2}}{v-r}}{\binom{\binom{n}{2}}{v}}.$$

The number of nodes of each degree in $G_{n.v}$ can be compared with the expected number under the Random Graph Hypothesis to test clustering tendency. When the number of nodes of high degree is larger than expected in a random graph, evidence for non-random data exists. One practical difficulty is choosing an appropriate value of v at which to apply the test.

The third test for clustering tendency computes the number of cycles of order k in an experimentally derived graph and compares it to that expected under the Random Graph Hypothesis, using an expression for the expected value.[59.p.105] The presence of clusters should force more cycles of order three or four than expected in a random graph.

(b) Ling[47] has proposed a means for testing the compactness of a clustering structure based on the number of nodes in clusters. In our framework, Ling's procedure provides information concerning clustering tendency. Suppose the number of nodes incident to one or more edges is observed in a threshold graph. Let $P(i; n, v)$ be the probability, under the Random Graph Hypothesis, that exactly i nodes are incident to some edge (i nodes are in connected sub-graphs of size 2 or more) in graph $G_{n.v}$. Ling[47] provides a recurrence relation for computing $P(i; n, v+1)$ given $P(i; n, v)$ for $i = 1, \ldots, n$. Define the cumulative distribution function as:

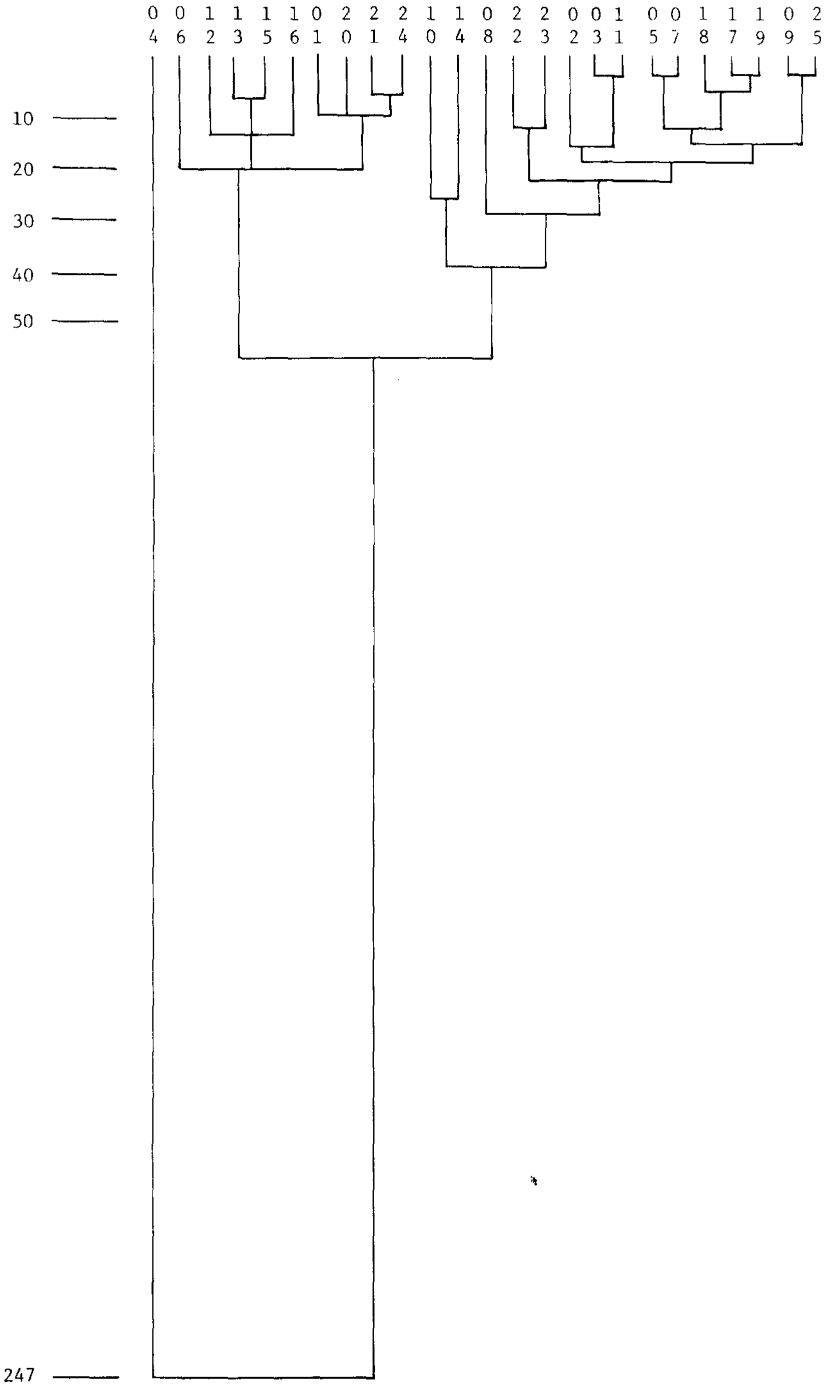

Fig. 6. Single link dendrogram (ordinal).

$$F(k;n,v) = \sum_{i=1}^{k} P(i;n,v).$$

This is the probability that at most k points are in non-singleton single-link clusters at level v under the Random Graph Hypothesis.

The observed number of points in clusters has cumulative distribution function $F(k;n,v)$ for given n and v under the Random Graph Hypothesis. The sequence of numbers $F(k;n,v)$ for $v = 1, 2, \ldots, n(n-1)/2$ thus indicates the levels, if any, at which k clustered points differ from that expected under the null hypothesis. Small values of $F(k^*; n, v)$, where k^* is the observed number of clustered points at level v, would lead to rejection of the null hypothesis.

The difficulty with this procedure is that its power is very poor. For example, if the data points were organized into a moderate number of compact clusters, all nodes would soon be in clusters and $F(k; n, v)$ would be large for all v. Also, the presence of a few

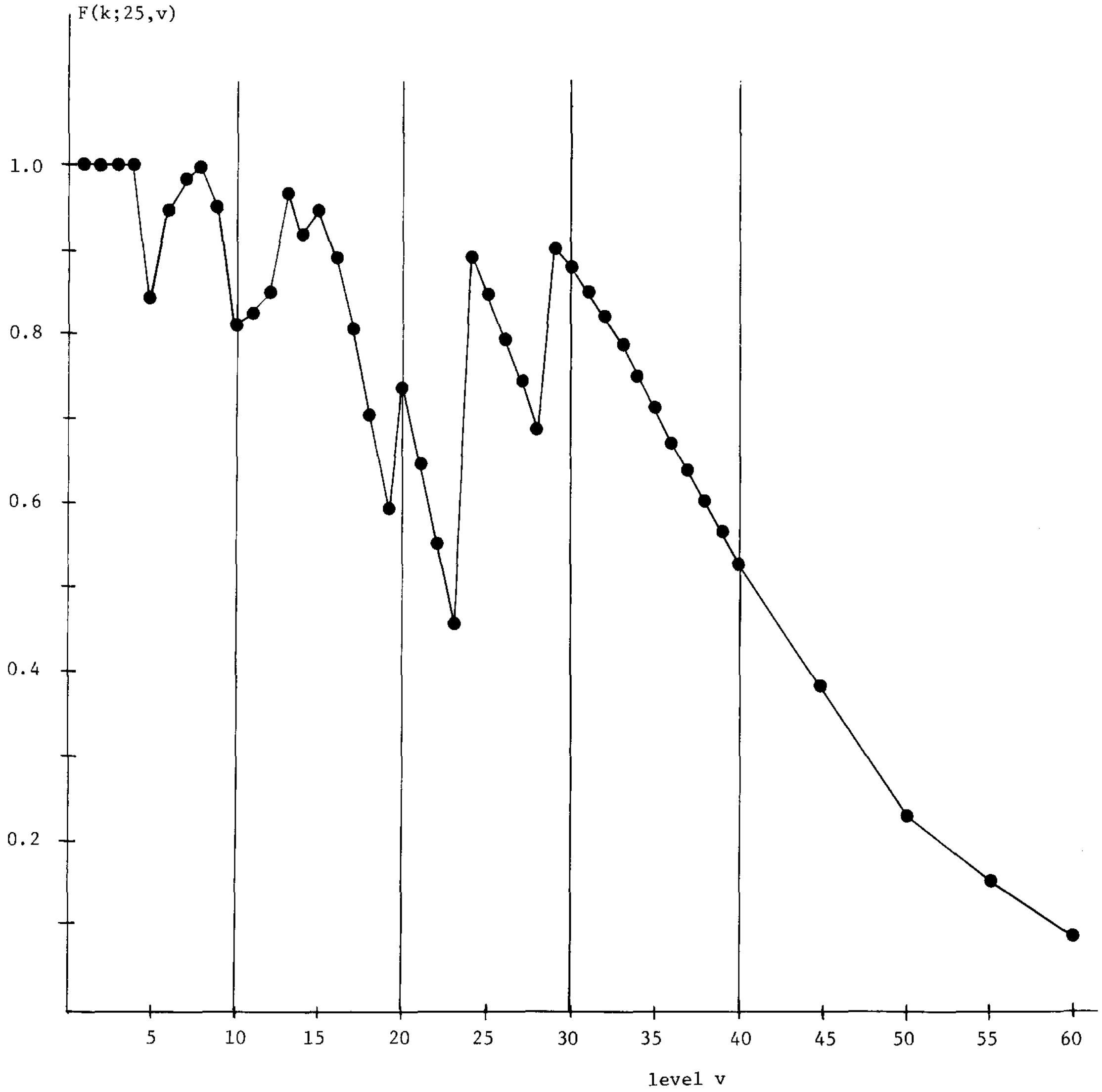

Fig. 7. Measure of clustering tendency based on number of clustered points.

outliers in otherwise random data would artificially delay the inclusion of all nodes in clusters, leading to unjustified rejection of the null hypothesis.

A plot of $F(k^*; 25, v)$ for the 25-sample choral speech data is shown in Fig. 7. The only small values occur for $v > 60$, due to the single outlier, sample 4. Thus, discounting the outlier, this procedure provides no evidence for concluding the data are non-random.

(c) Ling and Killough[50] suggest yet another test for clustering tendency based on the expected number of components in a random graph. A component in this context is either a single-link cluster or a single point. The number of components at level v is the number of intersections with a horizontal line across the single link dendrogram at level v. Ling and Killough provide tables for $E_{n,v}$, the expected number of components in an n-point graph at level v under the Random Graph Hypothesis. If the number of components observed at level v is significantly different from $E_{n,v}$, either higher or lower, one has evidence of a tendency to cluster.

Figure 8 is a plot of $E_{n,v}$ and the observed number of components for the 25-sample choral speech problem. The only variation of note occurs at large levels so there is no evidence of a tendency to cluster.

(d) Strauss[75] has suggested a means for testing clustering tendency in a pattern matrix. He characterized the clustering mechanism in terms of a "clustering parameter", V. When $V = 0$, no clustering exists so the test for clustering tendency consists of making a decision about the value of V. Kelly and Ripley[41] disproved the Strauss characterization lemma and provided a slightly different characterization. Strauss' work was continued by Saunders and Funk[65] who clarified several issues and demonstrated that, when the patterns are sufficiently sparse, the statistic used for testing clustering tendency has a known asymptotic distribution.

The test statistic is $Y_n(r)$, the number of interpoint distances that are r or less. That is, $Y_n(r)$ is the number of points captured in spheres of radius r centered at each of the n patterns in turn, not counting the points at the centers. Stated in rough terms, Saunders and Funk[65] showed that if n increases and the volume of the feature space increases as n^2, $Y_n(r)$ has, under the null hypothesis, an asymptotic Poisson distribution with parameter $\binom{n}{2}x$ [Vsphere/Vspace] where Vsphere is the volume of a sphere of radius r and Vspace is the volume of the space holding the patterns. The null hypothesis here is the Random Position Hypothesis which assumes a uniform distribution over the space. The test is to compare $Y_n(r)$ to a threshold chosen as the $(1-\alpha)$th percentile of the null distribution and reject the null hypothesis if $Y_n(r)$ is large.

Several practical difficulties must be overcome to implement this interesting idea. Determining Vspace can be difficult. Assuming, for example, a spherically-shaped volume can lead to inaccuracy if the volume is actually a hypercube and the inaccuracy worsens as d increases. If one rotates the feature space to uncorrelate the original features and applies a diagonal transformation to equalize variances, a spherical volume might be reasonable. One is then faced with choosing a suitable r. Saunders and Funk[65] suggest an intuitive test based on

$$G = \mathrm{SUP}\{Y_n(r)/E(Y_n(r))\},$$

the supremum being taken over a suitable range of r values. When d is large, say 10 or more, small changes in r, n, and the radius of the space can dramatically affect the results of the test. Saunders and Funk[65] require uniformly distributed points under the null hypothesis. If all the patterns had, say, a Gaussian distribution, the test would not be applicable.

We applied this test to the computer performance data (Section 1) and measured the clustering tendency. A linear transformation (based on eigenvectors) was applied to the data to make all features uncorrelated with mean zero and variance one. This allowed us to assume that the space was a hypersphere of radius three. Of the 300 patterns, 254 were retained in this space. Using the Saunders and Funk[65] asymptotic distribution for $Y_{254}(r)$, the null hypothesis was overwhelmingly rejected for values of r between 0.1 and 1.5. On the other hand, the null hypothesis was accepted for all values of r when the test was applied to artificially generated random data sets.

3.2 *Global fit of hierarchy*

We now consider the entire hierarchy produced by a hierarchical clustering method and ask whether it provides a valid conceptualization of a given proximity matrix. This question of global fit is crucial in applications so it is not surprising to see experience from the subject matter influencing judgment on the appropriateness of a hierarchy. We review some of the more common objective indices of global fit.

3.2.1 *Quantitative data.* Here we assume an interval scale for the proximities and try to evaluate the degree to which an entire hierarchy matches a proximity matrix. Sneath and Sokal[73] stress the importance of evaluating the match between the dendrogram and the proximity matrix and suggest the cophenetic correlation coefficient (CPCC) as a standard of comparison. Given a dendrogram whose level values are on the same scale as the entries in the $n \times n$ proximity matrix $D = [d_{ij}]$, an $n \times n$ cophenetic matrix $CP = [c_{ij}]$ is defined with c_{ij} being the first level in the dendrogram at which patterns i and j occur in the same cluster. The CPCC is the ordinary product-moment correlation coefficient between the entries of D and CP, viz:

$$\mathrm{CPCC} = \frac{(1/N)\Sigma\, d_{ij}c_{ij} - (\bar{d})(\bar{c})}{[(1/N)\Sigma\, d_{ij}^2 - \bar{d}^2]^{1/2}[(1/N)\Sigma\, c_{ij}^2 - \bar{c}^2]^{1/2}}$$

where $\bar{d} = (1/N)\Sigma\, d_{ij}$, $\bar{c} = (1/N)\Sigma\, c_{ij}$, $N = n(n-1)/2$, and all sums are for all values of i and j for which $i < j$. The CPCC is especially appropriate for interval-scale proximities. The larger the CPCC, the better the match between proximity matrix and dendrogram and the better the global fit.

By virtue of its definition, the entries of the cophenetic matrix satisfy the ultrametric inequality, viz:

$$c_{ik} \leq \text{maximum } (c_{ij}, c_{jk}), \quad \text{all } (i, j, k)$$

so the match between data and dendrogram cannot be perfect unless D is also ultrametric, a situation which seldom occurs in practice. One usually argues that the "true" underlying hierarchy is obscured by "noise" of some sort, thus masking the true nature of the data. The actual value of the CPCC depends to a great extent on the clustering method employed. For example, the UPGMA method[61] seems to produce consistently high values.

Rohlf and Fisher[62] studied the distribution of the CPCC under the hypothesis that the patterns are randomly chosen from a single distribution, either Gaussian or uniform and used the UPGMA clustering method. The average value of the CPCC tended to decrease with n and to be somewhat independent of the number of features for both distributions studied. The number of Monte Carlo runs was extremely small. The most surprising aspect of their study was that the CPCC was much more sensitive to the choice of correlation or distance as a proximity measure than to the underlying distribution of the patterns.

Rohlf and Fisher[62] exhibited approximate 5% thresholds for testing the null hypothesis of a single Gaussian cluster vs the alternative hypothesis of a system of nested clusters using the CPCC as a test statistic. A value of the CPCC above 0.8 seems to be sufficient evidence for rejecting the null hypothesis. However, Rohlf[61] warns that "... even a CPCC near 0.9 does not guarantee that the (dendrogram) serves as a sufficiently good summary of the phenetic relationships".

The tremendous number of applications of the CPCC, especially in numerical taxonomy, and familiarity with correlation coefficients have made the CPCC the most widely understood measure of global fit for quantitative data.

3.2.2 *Qualitative data.* When the proximity matrix contains data on an ordinal scale, a coefficient of rank correlation is used to measure the match between a dendrogram and the proximity matrix, in place of a CPCC. Hubert[33] proposed the Goodman–Kruskal γ-statistic for this purpose:

$$\gamma = \frac{S_+ - S_-}{S_+ + S_-}.$$

Here, S_+ is the number of "concordant", or consistent pairs while S_- is the number of "discordant" or inconsistent, pairs. Ties are not counted. A "pair" is a set of two duos of ranks. One duo involves proximity ranks and the other, phenetic ranks – corresponding to the same row–column positions. All $N(N-1)/2$ pairs must be checked, where $N = n(n-1)/2$. For example, suppose the $(1,2)$ and $(1,3)$ positions of the dissimilarity matrix contained ranks 3 and 8, respectively, while the $(1,2)$ and $(1,3)$ positions of the phenetic matrix contained values 5 and 12, respectively. The pair of duos $(3,8)$ and $(5,12)$ is concordant because $3 < 5$ and $8 < 12$. However, $(3,8)$ and $(6,7)$ would be discordant because $3 < 6$ but $8 > 7$. Finally, $(3,8)$ and $(3,12)$ are neither concordant nor discordant. The same is true for $(3,8)$ and $(5,8)$.

The γ statistic is invariant through monotone transformations on the proximity scale. The maximum value of γ is 1 and the minimum is -1. The larger γ, the better the match between dendrogram and proximity matrix.

Only when $\gamma = 1$ can one decide that the hierarchical structure imposed by a dendrogram truly "fits" the proximity matrix. In the case of a quantitative data scale, $\gamma = 1$ when the proximity matrix is ultrametric and either the single-link or complete-link methods (which give identical results for ultrametric priximity matrices) are used. How does one interpret other values of γ? Only in a comparative way. One can compare the fits imposed by two different clustering

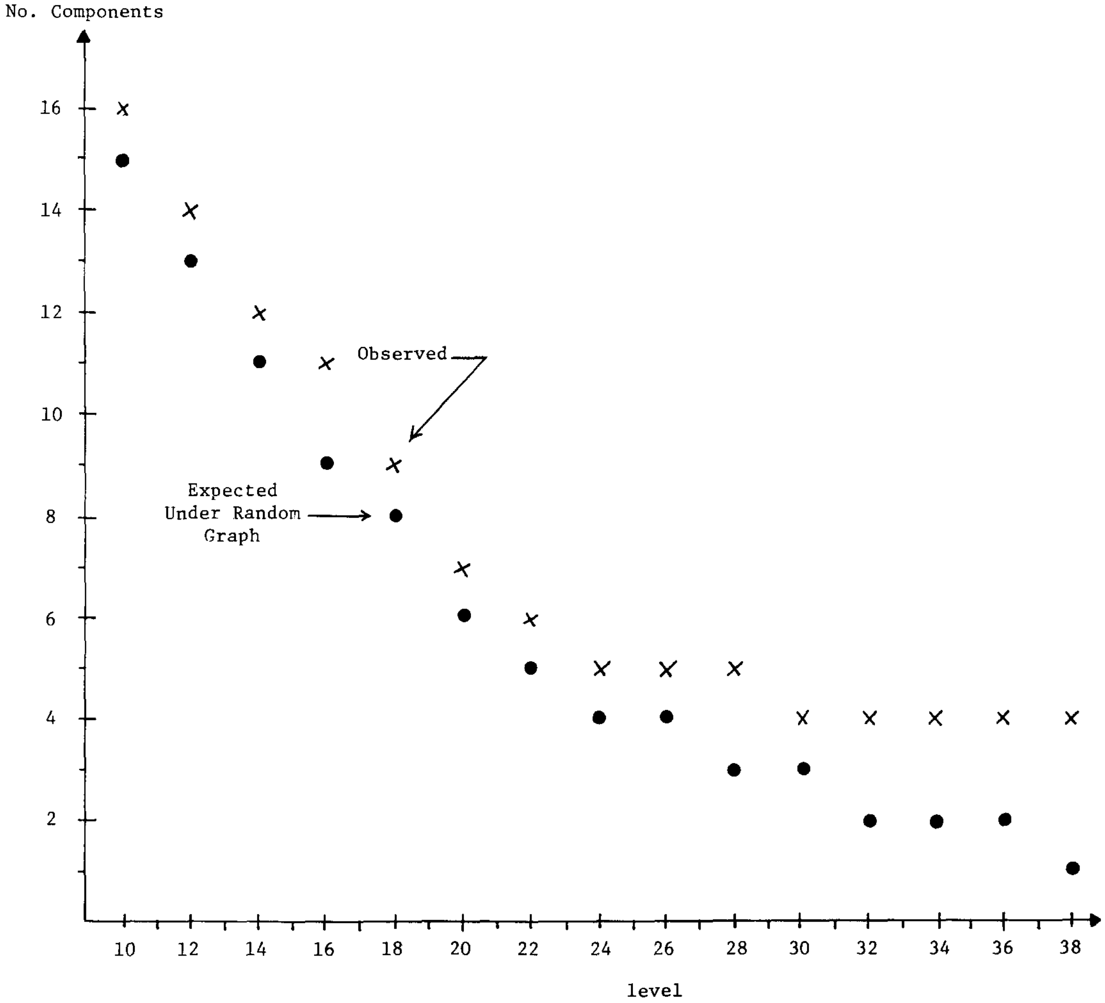

Fig. 8. Measure of clustering tendency based on number of components.

methods but cannot, in general, make absolute statements. Thus, in a strict sense, we are not dealing with a measure of global fit but with a theoretical tool for comparing clustering methods.

The value of γ in comparative studies of clustering methods was demonstrated by Baker[5] and Hubert.[33] Baker[5] reported a very interesting study that compared the single-link and complete-link methods and investigated the sensitivity of these methods to differences in the underlying clustering structure, based on γ. Specifically, Baker began with three "basal taxonomies" (chained, binary, and arbitrary) and generated proximity matrices by perturbing them with noise. The single-link method was expected to pick out the chained structure, the complete-link method was expected to be sensitive to the binary structure and the arbitrary structure was to provide an intermediate basis of comparison.

Baker's results showed that the mean γ for the complete-link method was better than for the single-link method in almost all situations. The fact that the complete-link method worked better even under a chained basal taxonomy is counterintuitive. Hubert[33] extended Baker's study and provided more evidence for the superiority of the complete-link method over the single-link method in matching certain dendrograms.

Hubert[33] emphasized the importance of sample size. He reported the results of Monte Carlo simulations in which, for fixed n, the distribution of γ under the Random Graph Hypothesis was estimated under complete-link and single-link clustering. Hubert[33] provides tables of percentage points of the estimated distribution function of γ for $4 \leq n \leq 16$ and tables of the mean and standard deviation of γ for $4 \leq n \leq 25$. He suggests that the quantity

$$n\gamma - a\ln(n)$$

has an approximate $N(0, 1)$ distribution where $a = 1.8$ under complete-link clustering and $a = 1.1$ under single-link.

The γ-statistics for the ordinal dissimilarity matrix derived in the 25-sample choral speech study were computed for the complete-link method (γ_{CL}) and the single-link method (γ_{SL}) with the following results.

$$\gamma_{CL} = 0.5368; \qquad \gamma_{SL} = 0.6853.$$

According to Hubert's simulations, the means (μ) and standard deviations (σ) for γ when $n = 25$ are:

$$\mu_{CL} = 0.23; \quad \mu_{SL} = 0.13; \quad \sigma_{CL} = \sigma_{SL} = 0.04.$$

Thus, γ_{CL} is over seven and γ_{SL} almost 14 standard deviations above the mean. This suggests that both hierarchies are appropriate for the data.

3.3 *Global fit criteria for partitions.*

Most clustering algorithms provide one or more clusterings or partitions of the data. Here we address the validity of the entire clustering, which may contain some clusters which are individually "real" and some which are not. The two questions of interest are: (i) What is an appropriate number of clusters? and (ii) How "real" is the clustering itself? The first question deals with trade-offs between parsimony and cluster homogeneity. The second compares a given clustering with random partitions of the same size. We concentrate on specific criteria and only mention in passing suggestions for studying the stability of clusterings, as suggested by Strauss *et al.*[76] and the sensitivity of clustering to small changes, as proposed by Strauss.[77]

3.3.1 *Partitions from a pattern matrix.* This section assumes that the n patterns are expressed in terms of d features in a pattern matrix and the features are all quantitative. The ideal clusters under this approach are compact, well separated, and have hyperellipsoidal shape. Clustering algorithms which partition a set of pattern vectors into groups or clusters are essentially based on minimizing a squared error criterion. Since the squared error is a monotonically decreasing function of the number of clusters, K, it is usually minimized for a fixed K. Thus, either the number of clusters desired must be specified *a priori* or the clustering method must determine a suitable value of K with the help of user-specified criteria for merging and splitting clusters.

The popularity of the minimum-squared-error partition methods is evident from the numerous algorithms which have appeared in the literature over the past 15 years.[2, 8, 18, 25, 27, 55, 69, 78] Even though these methods are computationally attractive and do not make any assumptions about the form of the densities of the subpopulations or groups present in the data, they do tend to define clusters that are hyperellipsoidal in shape. For this reason, the clustering methods which minimize a squared error criterion are sometimes characterized as nonparametric methods for fitting mixtures of Gaussian distributions to the data, also called unsupervised learning. Thus, before we can even talk about the validity of clusters obtained by minimum-squared-error partition methods, we should be willing to make an assumption that the patterns in each cluster are distributed approximately according to a multivariate Gaussian distribution.

A naïve approach to studying the validity of clusters obtained in a partition is to use the techniques of multivariate analysis of variance to test the null hypothesis that all the cluster centers are equal. The standard MANOVA test requires that the pattern vectors in each cluster have Gaussian distributions with a common covariance matrix. If the null hypothesis is rejected, pairwise comparison of the cluster centers can be made. Unfortunately, such tests are not valid because the labelling of patterns with cluster numbers is not done randomly. The clusters were formed to maximize the very inter-cluster distance that is tested in MANOVA.

Five specific studies of cluster validity with partitional clusters from a pattern matrix are summarized below.

(a) Choosing the proper value of K, the number of clusters in the partition, is a fundamental question in a study of cluster validity. An emprical approach used in answering this question is to plot the squared error for different values of K. If indeed there are K_0 clusters present in the data, we expect to see a "knee" in this curve at $K = K_0$. Duda and Hart[20] attempt to provide a statistical test to determine whether or not the decrease in the squared-error as a result of increasing the number of clusters from K to $(K+1)$ is significant. More specifically, they propose the null hypothesis that the given set X of n d-dimensional patterns is a random sample from a Gaussian distribution with mean vector μ and covariance matrix $\sigma^2 I$ where I is the identity matrix. Under the null hypothesis, the decrease in the mean-squared-error from splitting the n patterns into two groups would not be significant. Let $J_e(1)$ and $J_e(2)$ denote the squared-errors when the set X is treated as one cluster and when it is partitioned into subsets X_1 and X_2, respectively. Duda and Hart obtain the null distribution of $J_e(1)$ and $J_e(2)$ (for a suboptimal partitioning) and propose the following test: reject the null hypothesis at the p-per cent significance level if

$$\frac{J_e(2)}{J_e(1)} < 1 - \frac{2}{\pi d} - \alpha\sqrt{\left(\frac{2(1-8/\pi^2 d)}{nd}\right)},$$

where α is the $(1-p)$th percentile of the $N(0,1)$ distribution. The test can be applied to splitting of individual clusters in a K-cluster problem. It does not provide an answer to the basic question of the validity of the clusters.

Wolfe[84] suggested the likelihood ratio test for determining an appropriate value of K. To test the hypothesis of K components against $(K-1)$ components in the mixture, one computes

$$\chi^2 = -2\log\left(\frac{L_{K-1}}{L_K}\right),$$

where L_K is the likelihood of the sample based on K components. The quantity χ^2 has an asymptotic chi-square distribution with degree of freedom equal to the difference in the number of parameters under the hypothesis of K and of $(K-1)$ components. Thus the clustering should be repeated with $K=1, 2, \ldots,$ etc. until the observed χ^2 value is below some critical threshold.

Hall *et al.*[30] take a similar approach except that they test the validity of a cluster by a multivariate version of the Kolmogorov–Smirnov (K–S) test. First, they test the given set of data for unimodality by applying the multivariate version of the K–S test. If the test fails, they split the data into two clusters by using the basic ISODATA procedure; otherwise the program stops. Each cluster, in turn, is split unless it passes the K–S test. Thus the goal is to obtain the smallest number of multivariate Gaussian clusters. As pointed out by Hall,[30] this test is limited to data in low dimensions due to the exponential growth of the computation (n^{d+1}).

(b) Hartigan[31] investigates a measure of splitting the data that is related to cluster validity. Consider first the one-feature problem when the n patterns are split so as to minimize SSW, the within-cluster sum of squares. Hartigan assumes that cluster C_i is distributed as $N(\mu_i, \sigma^2)$, $i=1,2$. The null hypothesis is: $\mu_1 = \mu_2$. Engleman and Hartigan[21] obtain the distribution of (SSB/SSW) under the null hypothesis when the patterns are partitioned into two clusters based on the minimization of SSW, when SSB is the between-cluster sum of squares. In one dimension, only the $(n-1)$ possible splits need be considered, as opposed to the total number of partitions of n patterns into two clusters. Hartigan[31] demonstrates that log (SS-B/SSW) has an asymptotic Gaussian distribution. If the observed value of log (SSB/SSW) is large compared to the expected value at some significance level then the null hypothesis will be rejected at that level.

The above test was extended by Hartigan to clusters in multivariate space and the null hypothesis is that all cluster centers are equal. Under this null hypothesis, the patterns belong to a multivariate Gaussian distribution. The expected value and the variance of (SS-B/SSW) for large n and d are known. Again, if the observed value of (SSB/SSW) is several standard deviations away from its expected value then the null hypothesis is rejected.

(c) Sneath[72] has proposed an interesting test for the distinctiveness of clusters in terms of an index of overlap, V_G, and an index of disjunction, W, which are obtained from the projection of cluster members onto the intercentroid axis of the two clusters. In determining his test statistic, t_w, Sneath assumes that the clusters are approximately hyperspherical and their members are distributed according to multivariate Gaussian distributions.

His test statistic can be summarized as follows: if we have two clusters C_1 and C_2 containing n_1 and n_2 patterns in a hyperspace with intercentroid distance Δ, then the patterns belonging to the two clusters can be projected on the intercentroid axis. These projections define two random variables, Q_1 and Q_2, with sample standard deviations s_1 and s_2. Sneath's index of overlap, V_G, is one when $\Delta = 0$ and zero when $\Delta = \infty$. The index of disjunction, W, is defined as

$$W = \Delta/[(n_1 + n_2)(s_1^2/n_1 + s_2^2/n_2)]^{1/2}.$$

The quantity

$$t_w = W\sqrt{(n_1 + n_2)}$$

has a non-central t-distribution under the null hypothesis.

This statistic is used to determine whether the observed disjunction is significantly greater than a predetermined value. The test provided by Sneath is conservative in the sense that small values of disjunction or, alternatively, large values of overlap do not necessarily mean that the clusters are not distinct in the multidimensional space.

In order to determine a critical value of W, $W_{\exp}$, and correspondingly a value $t_{w_{\exp}}$ for t_w, Sneath considers the null hypothesis that the two clusters are formed by splitting an essentially uniform swarm of points. More specifically, his null distribution is in the shape of a cylinder (length longer than the diameter) in a hyperspace containing points according to a Poisson distribution. This particular form of the null hypothesis was chosen because the expected value of the overlap can be readily obtained, $W_{\exp} = \sqrt{3}$. A value for $W_{\exp}$ of $\sqrt{3}$ corresponds to $V_G = 8.326\%$. To reject the null hypothesis, W_{obs} and $V_{G_{\mathrm{obs}}}$ must be significantly greater than $\sqrt{3}$ and less than 8.326, respectively.

(d) Another theoretical model for clustering, which is related to the approach based on minimizing a squared error criterion as discussed earlier, is to assume that the patterns are drawn from a finite mixture of multivariate distributions. The mixture density can be denoted as

$$f(x) = \sum_{i=1}^{K} P(w_i) f(x|w_i),$$

where K is the number of components, $P(w_i)$ is a mixing parameter and $f(x|w_i)$ is a component density. It is generally assumed[30,67,68,84] that the component density, $f(\cdot|w_i)$, is multivariate Gaussian with unknown parameters μ_i, Σ_i, $i = 1, \ldots, K$. The clustering problem then reduces to estimating the parameters μ_i, Σ_i and $P(w_i)$, $i = 1, \ldots, K$. The usual approach is to obtain the maximum-likelihood estimates of μ_i, Σ_i and $P(w_i)$ based on the available samples. Once these parameters are estimated, patterns are assigned to the closest cluster or component using an appropriate metric and the tests proposed by Hartigan[31] and Sneath[72] can be used to test the distinctness of the clusters.

(e) Mountford[54] assumes a mathematical model for the n^2 entries in the proximity matrix D, rather than for the patterns themselves, viewing them as jointly distributed, Gaussian random variables based on the following equation:

$$d_{ij} = \mu_{ij} + g_i + g_j + l_{ij}.$$

Here, μ_{ij} is the nominal value of the proximity, g_i and g_j are effects peculiar to pattern i and j, and l_{ij} is the effect of interaction between patterns i and j. The analogy to an analysis of variance model based on the entries in a pattern matrix is clear. Mountford's proximities are similarities.

The covariance between d_{ij} and d_{kl} depends on the number of matches in the subscripts. Thus, d_{ij} and d_{kl} will be independent if $i, j \neq k, l$, but the indices d_{ij} and d_{ik} will be correlated.

Only the two-cluster case is considered. The ideal clustering has proximities for all pairs of patterns within the same cluster (μ_{11} and μ_{22}) which are greater than those for all pairs of patterns from separate clusters (μ_{12}). The natural measure of separation between the two clusters is $\mu_{11} + \mu_{22} - 2\mu_{12}$. The larger this measure, the more "real" the clusters.

The null hypothesis is: $\mu_{11} = \mu_{22} = \mu_{12}$ which corresponds to "no clustering". Any clusters generated by a clustering method under H_0 would be strictly artifacts of the clustering method. Mountford suggests a simplified statistic, denoted "b", to test this hypothesis.

The distribution of b under H_0 is very difficult to obtain because the sample has been treated by a clustering algorithm. Mountford suggests a conceptual strategy for approximating the distribution of an upper bound on b.

3.3.2 *Partitions from a hierarchy.* In hierarchical clustering, partitions are achieved by cutting a dendrogram or selecting one of the clusterings in the nested sequence of clusterings that comprises the hierarchy. Several such clusterings are available so it is essential, in applications, to be able to select those clusterings which have objective merit and discard those which are artifacts of the clustering method.

(a) Baker and Hubert[7] have studied this problem of the adequacy of a partition for an ordinal proximity matrix. Their results apply only to a very special case. However, their work demonstrates the inherent difficulty in this problem of validating partitions.

To begin, a characteristic function for the partition at level m in a hierarchy is defined as:

$$\begin{aligned} T_m(x_i, x_j) &= 0 \text{ if patterns } x_i \text{ and } x_j \text{ are in the same cluster} \\ &= 1 \text{ else.} \end{aligned}$$

If $r(x_i, x_j)$ is the (given) rank of the proximity between the pair (x_i, x_j) among all $n(n-1)/2$ possible pairs, where the more similar the pair the lower the rank in numerical value, then a measure of rank correlation between these ranks and the values assigned by the characteristic function T_m should measure a degree of correspondence. Baker and Hubert[7] suggest the Goodman-Kruskal γ-statistic, discussed in Section 3.2.2, as an appropriate measure of rank correlation. Since the partition in question is at level m, we denote the statistic as γ_m in this section.

Ideally, the proximity ranks $r(x_i, x_j)$ for all pairs within the same cluster should be less than the proximity ranks for pairs containing patterns in different clusters. Thus, T_m reflects the "ideal" or "true" clustering structure imposed by the partition and γ_m allows comparison to the actual data structure in the proximity matrix. Since only ranks are used, this method is essentially restricted to ordinal proximity matrices.

Baker and Hubert[7] try to phrase the problem of partitional adequacy in a hypothesis testing framework. The null hypotheses must reflect a no-structure or randomness condition. The most evident statement of randomness is the Random Graph Hypothesis. The types of partitionings achieved depend on the clustering method. The number of clusters in the partition is an important factor. To manage all these things, Baker and Hubert use a threshold dendrogram (so there are exactly $n-1$ partitions in every hierarchy), fix the number of patterns at 12, and fix the level at $m = 9$. The distribution of γ_9 under the null hypothesis has not yet been determined theoretically, although not for lack of effort, so Monte-Carlo simulations are necessary.

The null hypothesis is only half the story. Deciding that a clustering is not random is one thing, but deciding that it has a certain structure is quite another. How to form a meaningful alternative hypothesis? This alternative must be composed *before* the data are clustered and a particular clustering is achieved at level 9. Such an alternative hypothesis can be expressed by choosing a matrix, T, to reflect the ideal structure under this alternative. Baker and Hubert estimate the distribution of γ_9 under several alternatives.

The Baker and Hubert study also permits a comparison of the two clustering methods. The complete-link method was more powerful than the single-link for ideal partitions containing fewer than seven patterns within the larger class. The answers cannot be generalized to other cases. The methodology employed would need to be repeated for other values of n and m.

(b) A second approach to validating partitional structures called "quadratic assignment" has been proposed by Hubert and Schultz[38] as part of a very large study of schemes for validating data structures. We begin with an *a priori* notion of an ideal structure which specifies not only the number of clusters, but also the number of patterns in each cluster. This ideal structure is reflected in a structure matrix $[C(i,j)]$. Only the blocks along the main diagonal contain non-zero elements. The ith main-diagonal blocks contain zeros on their main diagonals and the number k_{ii} elsewhere, where $k_{ii} = 1/n_i(n_i - 1)$ and n_i is the number of patterns in cluster i.

Letting $D = [d_{ij}]$ be a dissimilarity matrix, the following statistic is used to test the partition.

$$\Gamma(g) = \sum_{i,j} d_{ij} C[g(i), g(j)]$$

where g indicates a permutation function on the integers from 1 to n and the sum is over all n values of i and of j. This statistic is a product moment between the given dissimilarities and the ideal structure after the rows and columns of C have been rearranged according to g.

If g_0 is the identity permutation, so $g_0(k) = k, k = 1, \ldots, n$, then $\Gamma(g_0)$ is the sum of the average within-cluster dissimilarities, summed over all clusters. The smaller $\Gamma(g_0)$, the better the partition defined by g_0 fits the data.

The idea behind the validity test is to pick out "unusual" clusterings of type $(n_1, n_2, \ldots, n_K)$ where K is the number of clusters and $n_1 + n_2 + n_K = n$. Conceptually, one would compute $\Gamma(g)$ for all $n!$ permutations of the pattern numbers and develop the sampling distribution of $\Gamma(g)$. Then, one could see if a particular permutation, say g_0, were unusual by seeing where $\Gamma(g_0)$ occurs in the distribution. One can think of the sample distribution created in this way as the distribution of $\Gamma(g)$ under the null hypothesis that all permutations are equally likely.

The unique aspect of this approach is that formulas for the mean, $E(\Gamma)$ and variance, $\mathrm{Var}(\Gamma)$, of Γ are known.[38] Thus, the entire distribution of $\Gamma(g)$ need not be estimated by Monte-Carlo simulations as was necessary in a related study by McClain and Rao.[52] Instead, one can compute the statistic:

$$\frac{\Gamma(g_0) - E(\Gamma)}{[\mathrm{Var}(\Gamma)]^{1/2}}$$

to determine how large is "large".

(c) Baker and Hubert[6] begin with a proximity matrix on an ordinal scale and use the notion of an isolated maximal complete subgraph (isolated clique) as the ideal cluster. The complete-link clustering method generates cliques and their work concerns the validity of complete-link clusters. The perfect clustering can be pictured as a threshold graph having each node involved in a complete subgraph with no edges between the subgraphs.

The number of "extra" edges in the threshold graph under the Random Graph Hypothesis is the index used to judge the validity of a clustering. Let T_k be the minimum number of edges required in the threshold graph G_c to define this structure. If A_k is the number of edges in G_c, the number of extraneous edges is

$$E_k = A_k - T_k.$$

The larger the value of E_k, the worse the fit of the complete-link structure to the partition achieved by the complete-link clustering method.

Sample size introduced a practical difficulty. The distribution of E_k is highly dependent on n. Thus, n must be fixed in a Monte-Carlo procedure. Baker and Hubert[6] used 500 Monte-Carlo trials and produced tables of the joint distribution of A_k and T_k for several k and for $n = 8, 12$, and 16. One can test the validity of the clusters generated by a complete-link clustering method when n is 8, 12, or 16 by comparing the observed value of E_k to a distribution of E_k computed from the table.

3.4 *Validity of individual clusters*

We now review criteria which measure the degree to which an individual cluster can be termed "real". Measures of compactness and isolation are discussed for clusters, relative to their environment.

3.4.1 *Isolation criteria in a hierarchy.* A cluster is "real" if it forms early in the dendrogram for its size and lasts a relatively long time before being swallowed up. Ling[47,49] suggests measuring the compactness of a cluster by its birth size and measuring the isolation of an individual cluster by the cluster's lifetime. Ling works under the Random Graph Hypothesis. Thus, we will consider dendrograms in which the birth level of a cluster is the number of the threshold graph at which the cluster first forms.

If c_1 is the birth level of cluster C and c_2 is the lowest level at which C becomes a proper subset of another cluster, the isolation index for C is defined as:

$$I(C) = c_2 - c_1.$$

Several other indices could be considered as isolation indices. For instance, the number of "extra" edges in a threshold graph, Section 3.3.2(c), measures the isolation of a single-link clustering. The number of inter-cluster dissimilarities smaller than the intra-cluster dissimilarities indicates the degree to which a cluster is isolated from other clusters in a partition. Here, we consider only the lifetime, or survival "time" of a cluster in a hierarchy.

Ling[47] determined the exact distribution for $I(C)$ under the Random Graph Hypothesis and provided a recurrence equation for computing the probability mass function using single-link clustering along with expressions for the mean and variance. Thus, the probability that the lifetime, $I(C)$, of a cluster of size $N(C)$ exceeds a value i, denoted

$$F(i) = \text{Prob}[I(c) > i \,|\, N(C)]$$

can be determined under the Random Graph Hypothesis.

Table 1 shows $F(i)$ for all the single-link clusters in the 25-sample choral speech data. The clusters are identified by the birth levels in Fig. 6. The significant clusters, according to this test, are marked by asterisks. Cluster C_{20} in Table 1 is cluster C_F in Fig. 1 and, by this measure, is a well-isolated cluster. However, clusters C_A and C_B in Fig. 1, born at levels 2 and 5, respectively, are not significantly isolated. Clusters which are not significant under the Random Graph Hypothesis are probably chance clusters whereas "significant" clusters must be tested further.

Hartigan[31] has proposed a procedure for studying the reality of clusters that depends not only on the clustering method but also on the algorithm chosen to implement the method. Several examples of clustering criteria for which no sampling theory has been developed are provided in the literature.[24, 83]

3.4.2 *Compactness criteria in a partition.* We now review measures of the compactness of individual clusters from a partition. Any global fit criterion for partitions, Section 3.3.2, that is applied for two clusters can be made to measure compactness.

One other approach to validating an individual cluster is to fashion a structure matrix reflecting perfect structure and apply the Hubert and Schultz[38] Γ statistic discussed in Section 3.3.2. Consider a cluster of n_1 patterns and let $n_2 = n - n_1$ patterns be outside the cluster. The validity of the two-cluster clustering was considered in Section 3.3.2. Here, we are interested only in one cluster. An intuitively appealing statistic is:

(average of within-cluster dissimilarities) $-$ (average of dissimilarities between the cluster and outside).

A structure matrix that makes the statistic equal to this difference is indicated below.

$$\varepsilon = \begin{array}{r} \\ n_1\{ \\ \\ n_2\{ \end{array} \begin{array}{c} \overbrace{\phantom{k_{11}}}^{n_1} \quad \overbrace{\phantom{-k_{12}}}^{n_2} \\ \left[\begin{array}{c:c} k_{11} & -k_{12} \\ \hdashline -k_{12} & 0 \end{array}\right] \end{array}$$

The entries in the $n_1 \times n_1$ submatrix are all k_{11}, except for zero diagonal entries, where $k_{11} = 1/(n_1(n_1-1))$. The entries in the $n_1 \times n_2$ and $n_2 \times n_1$ submatrices are all $-k_{12}$, where $k_{12} = 1/(2n_1 n_2)$. The $n_2 \times n_2$ submatrix contains all zeros. If g is any permutation of the integers $(1, \ldots, n)$,

$$\Gamma(g) = \sum_{i,j} d_{ij} C[g(i), g(j)]$$

is the difference cited above when patterns $g(1)$, $g(2)$, $\ldots$, $g(n_1)$ are in the cluster.

This statistic is used as explained in Section 3.2.2. The number of patterns in the cluster must be fixed *a priori*. The null hypothesis is that all permutations are equally likely. Permutations producing very small values of Γ are unlikely and are taken as evidence for a valid cluster. Hubert and Schultz provide expressions for the mean and variance of Γ.

4. CONCLUSIONS AND SUMMARY

This paper provides a framework for classifying all approaches to the problem of verifying the clusters obtained from the most common clustering methods. The factors most directly affecting the cluster validity

Table 1. Lifetimes of single-link clusters for 25-sample choral speech data

Cluster	Birth level	Size	Lifetime i	$F(i)$
C_1: (17, 19)	1	2	4	0.4893
C_2: (5, 7)	2	2	12	0.8718
C_3: (3, 11)	3	2	12	0.8728
C_4: (9, 25)	4	2	13	0.8942
C_3: C_1, (18)	5	3	9	0.9013
C_6: (21, 24)	6	2	4	0.4956
C_7: (13, 15)	7	2	4	0.4969
C_8: (1, 20)	8	2	2	0.2907
C_{10}: C_8, C_6	10	4	8	0.9377
C_{11}: C_7, (12)	11	3	2	0.4052
C_{12}: C_{11}, (16)	12	4	6	0.8764
C_{13}: (22, 23)	13	2	9	0.7974
C_{14}: C_2, C_5	14	5	3	0.7265
C_{15}: C_3, (2)	15	3	4	0.6536
C_{17}: C_{14}, C_4	17	7	2	0.6931
C_{18}: C_{10}, C_{12}	18	8	2	0.7328
C_{19}: C_{15}, C_{17}	19	10	3	0.8999
C_{20}: C_{18}, (6)	20	9	34	*
C_{22}: C_{19}, C_{13}	22	12	7	*
C_{24}: (10, 14)	24	2	16	0.9508
C_{29}: C_{22}, (8)	29	13	11	*
C_{40}: C_{24}, C_{29}	40	15	14	*
C_{52}: C_{20}, C_{40}	52	24	195	*

* $F(i) > 0.99$.

problem are listed to provide the user of clustering algorithms a comprehensive picture of the entire problem. Clustering methods themselves are not discussed. The main result of this paper is a compendium of proposed solutions to the cluster validity problem and a list of caveats, especially for the uninitiated.

Specific methodologies for verifying clusters and clusterings in an absolute sense do not exist for any reasonably large class of clustering problems. Many more results and tests are available for hierarchical clusters than for clusters of points in a space, which indicates that cluster validity has received more attention in psychometric than in engineering applications. Certain specific situations have been thoroughly studied, such as the level-nine clustering in a complete-link hierarchy on twelve items described by an ordinal proximity matrix.(7) Unfortunately, the results cannot be generalized to other sample sizes. Such studies illustrate the difficulty of the basic problem. The choral speech data we considered in this paper shows the limitations of the current state-of-the-art in cluster validity. Without considering the outlier (sample no. 4), the data show no tendency to cluster but the individual single-link clusters C_M and C_F turn out to be significant and interpretable as groups of male and female speakers. In addition, the single-link and complete-link hierarchies fit the data well. All of these results are based on the Random Graph Hypothesis whose use is questionable because the original proximity matrix was on a ratio scale.

One should not expect to find a single statistic to serve as a panacea for all problems in cluster validity. Too little prior information is available in the typical clustering problem and too many factors are involved to expect a single statistic to cover the validity of clusters even for a single class of problems, such as partitions of points in a metric space. After all, cluster analysis itself is exploratory in nature and a "tool for discovery"(2) rather than an end in itself. Even when one is willing to make the usual assumption of multivariate statistics, such as Gaussian distributions for each cluster and equality of scatter in all clusters, the standard "t" and "F" tests are, at best, useless and, at worst, misleading. The "t" and "F" statistics themselves can be computed but their distributions under "randomness" null hypotheses are unknown in a clustering situation. That is, the cluster labels were placed on the patterns to optimize the very criterion that these statistics are supposed to test. This seemingly obvious fact has been repeatedly ignored in the literature.(26, 70, 80) Since one is forced into Monte Carlo studies(5, 6, 7, 10, 42, 54, 72, 76) to approximate null distributions, it makes sense computationally to adopt simpler statistics than the ones normally used in multivariate statistics.

A user of clustering algorithms interested in cluster validity would be well advised at present to apply several clustering approaches and check for common clusters instead of searching for a technical measure of validity for an individual clustering. At present, the application of category information and knowledge from the subject matter area about what "makes sense" is a more fruitful endeavor than applying criteria that depend only on the data unless one is willing to engage in large-scale Monte Carlo studies that will apply only to the specific problem in hand.

A few studies comparing the merits of clustering methods and clustering algorithms are available,(10, 15, 16, 18, 33, 45, 58) but little agreement exists on appropriate criteria. The measures of cluster validity and some of the prior structures discussed in this paper can be used to compare the efficacy of clustering algorithms. Some of our earlier work(18) has shown that seemingly disparate algorithms actually produce the same results and that close inspection of the methods themselves reveals why this is so. We hope that more efforts will be made to establish theoretical-based principles for comparing existing clustering algorithms and evaluating new ones.

We do not wish to be unduly pessimistic about the state of the art in cluster validity but we have tried to be realistic. The problem is difficult, mainly unsolved, and full of traps for the unwary user.(89, 90) Clearly, vigorous research programs are needed to develop effective methodologies for cluster validity.

Acknowledgement – The authors wish to thank Thomas Bailey for many helpful discussions and suggestions during the preparation of this paper.

5. REFERENCES

1. A. K. Agrawala, J. M. Mohr and R. M. Bryant, An approach to the workload characterization problem, *Computer* **9**, 18–32 (1976).
2. M. R. Anderberg, *Cluster Analysis for Applications.* Academic Press, N.Y. (1973).
3. P. Arabie, Clustering representations of group overlap, *J. math. Soc.* **5**, 113–128 (1977).
4. E. Backer, *Cluster Analysis by Optimal Decompositon of Induced Fuzzy Sets.* Delft University Press, Delft, Holland (1978).
5. F. B. Baker, Stability of two hierarchical grouping techniques, case 1 – sensitivity to data errors, *J. Am. statist. Ass.* **69**, 440–445 (1974).
6. F. B. Baker and L. J. Hubert, A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering, *J. Am. statist. Ass.* **71**, 870–878 (1976).
7. F. B. Baker and L. J. Hubert, Measuring the power of hierarchical cluster analysis, *J. Am. statist. Ass.* **70**, 31–38 (1975).
8. G. H. Ball, Data analysis in the social sciences – what about the details?, *Proc. Fall. Joint Comp. Conf.* 533–554 (1965).
9. J. C. Bezdek, Cluster validity with fuzzy sets, *J. Cybernet*, **3**, 58–73 (1974).
10. R. K. Blashfield, Mixture model tests of cluster analysis – accuracy of four agglomerative hierarchical methods, *Psychol. Bull.* **83**, 377–388 (1976).
11. R. E. Bonner, On some clustering techniques, *IBM Journal*, 22–32 (1964).
12. C. Bordone, R. Pisani, G. Sacerdote, R. Dubes and O. Tosi, Invariance of Talker's choral spectra, *J. Acous. Soc. Am.* **55**, S43 (1974).
13. W. J. Borucki, D. H. Card and G. C. Lyle, A method of using cluster analysis to study statistical dependence in

multivariate data, *IEEE Trans. Comput.* **C-24,** 1183–1191 (1975).

14. A. J. Cole, *Numerical Taxonomy*. Academic Press, New York (1969).
15. R. M. Cormack, A Review of classification, *J. R. statist. Soc.* Series A, **134,** 321–367 (1971).
16. K. M. Cunningham and J. C. Ogilvie, Evaluation of hierarchical grouping techniques – A Preliminary Study, *Comput. J.*, **15,** 209–213 (1972).
17. W. H. E. Day, Validity of clusters formed by graph–theoretic cluster methods, *Math. Biosciences*, **36,** 299–317 (1977).
18. R. Dubes and A. K. Jain, Clustering techniques – the users dilemma, *Pattern Recognition*, **8,** 247–260 (1976).
19. R. Dubes and A. K. Jain, Models and methods in cluster validity, *Proc. IEEE Conf. on Pattn. Recogn. and Image Process.*, Chicago, 148–155 (1978).
20. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York (1973).
21. L. Engelman and J. A. Hartigan, Percentage points of a test for clusters, *J. Am. Statist. Ass.* **64,** 1647–1648 (1969).
22. P. Erdös and A. Rényi, On the evolution of random graphs, *Mathl. Inst. Hung. Acad. Sci.* **6,** 17–61 (1960).
23. P. Erdös and A. Rényi, On Random Graphs – 1, *Publ. math. Debrecen*, **6,** 290–297 (1959).
24. G. F. Estabrook, A mathematical model in graph theory for biological classification, *J. Theor. Biol.* **12,** 297–310 (1966).
25. B. Everitt, *Cluster Analysis*. John Wiley, New York (1974).
26. S. Fillenbaum and A. Rapoport, *Structures in the Subjective Lexicon*. Academic Press, New York (1971).
27. H. P. Friedman and J. Rubin, On some invariant criteria for grouping data, *J. Am. Statist. Ass.* **62,** 1159–1178 (1967).
28. H. P. Friedman and J. Rubin, The logic of the statistical method, *The Borderline Syndrome*, R. Grinker, B. Werble and R. Drye, eds., Chapter 5. Basic Books, New York (1968).
29. I. Gitman and M. D. Levine, An algorithm for detecting unimodal fuzzy sets and its application as a clustering Technique, *IEEE Trans. Comput.* **C-19,** 583–593 (1970).
30. D. J. Hall, R. O. Duda, D. A. Huffman and E. E. Wolf, Development of new pattern recognition methods, Aerospace Research Laboratories, AD-7726141 (1973).
31. J. A. Hartigan, *Clustering Algorithms*, John Wiley, New York (1975).
32. E. G. Henrichon, Jr. and K. S. Fu, A nonparametric partitioning procedure for pattern classification, *IEEE Trans. Comput.* **C-18,** 614–624 (1969).
33. L. Hubert, Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures, *J. Am. Statist. Ass.* **69,** 698–704 (1974).
34. L. J. Hubert, Some applications of graph theory to clustering, *Psychometrika*, **39,** 283–309 (1974).
35. L. Hubert, Min and max hierarchical clustering using asymmetric similarity measures, *Psychometrika* **38,** 63–72 (1973).
36. L. Hubert, Monotone invariant clustering procedures, *Psychometrika* **38,** 47–62 (1973).
37. L. Hubert, Some extensions of Johnson's hierarchical clustering algorithms, *Psychometrika* **37,** 261–274 (1972).
38. L. Hubert and J. Schultz, Quadratic assignment as a general data-analysis strategy, *Br. J. math. statist. Psychol.* **29,** 190–241 (1976).
39. N. Jardine and R. Sibson, *Numerical Taxonomy*. John Wiley, New York (1971).
40. R. A. Jarvis and E. A. Patrick, Clustering using a similarity measure based on shared near neighbors, *IEEE Trans. Comput.* **C-22,** 1025–1034 (1973).
41. F. P. Kelly and B. D. Ripley, A note on Strauss's model for clustering, *Biometrika* **63,** 357–360 (1976).
42. M. G. Kendall, The basic problems of cluster analysis, *Discriminant Analysis and Applications*, T. Cacoullos, ed., pp. 179–191. Academic Press, New York (1973).
43. J. Kittler, A locally sensitive method for cluster analysis, *Pattern Recognition* **8,** 23–33 (1976).
44. W. L. Koontz, P. M. Narendra and K. Fukunaga, A graph-theoretic approach to nonparametric cluster analysis, *IEEE Trans. Comput.* **C-25,** 936–944 (1976).
45. F. K. Kuiper, A monte carlo comparison of six clustering procedures, M. S. Thesis, University of Washington (1971).
46. R. F. Ling, An exact probability distribution on the connectivity of random graphs, *J. math. Psychol.* **12,** 90–98 (1975).
47. R. F. Ling, Probability theory of cluster analysis, *J. Am. Statist. Ass.* **68,** 159–164 (1973).
48. R. F. Ling, The expected number of components in random linear graphs, *Ann. Prob.* **1,** 876–881 (1973).
49. R. F. Ling, Theory and construction of *K*-clusters, *Comput. J.* **15,** 326–332 (1972).
50. R. F. Ling and G. S. Killough, Probability tables for cluster analysis based on a theory of random graphs. *J. Am. Statist. Ass.* **71,** 293–300 (1976).
51. D. W. Matula, The largest clique size in a random graph, Technical Report CS7608, Dept. Comp. Sci., Southern Methodist Univ. (1976).
52. J. O. McClain and V. R. Rao, *CLUSTSIZ*: A program to test for the quality of clustering of a set of objects, *J. Marketing Res.* **12,** 456-460 (1975).
53. L. L. McQuitty, A mutual development of some typological theories and pattern analytic methods, *Ed. Psychol. Measur.* **27,** 21–48 (1967).
54. M. D. Mountford, A test for the difference between clusters, *Statistical Ecology*, G. P. Patil *et al.*, eds., Vol. 3, pp. 237–251. Penn. State Univ. Press, University Park, Pa. (1970).
55. A. N. Mucciardi and E. E. Gose, An automatic clustering algorithm and its properties in high-dimensional spaces, *IEEE Trans. Syst. Man Cybernet.* **SMC-2,** 247–254 (1972).
56. J. D. Orford, Implementation of criteria for partitioning a dendrogram, *Math. Geol.* **8,** 75–85 (1976).
57. K. Pettis, T. Bailey, A. K. Jain and R. Dubes, An intrinsic dimensionality estimator for near-neighbor information, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-1, 25–37 (1979).
58. W. M. Rand, Objective criteria for the evaluation of clustering methods, *J. Am. Statist. Ass.* **66,** 846–850 (1971).
59. A. Rapoport and S. Fillenbaum, An experimental study of semantic structures, *Multidimensional Scaling*, A. K. Romney, R. N. Shepard and S. B. Nerlove, eds, Vol. 2, pp. 93–131. Seminar Press, New York (1972).
60. R. J. Riddell and G. E. Uhlenbeck, On the theory of the virial development of the equation of state of monotomic gases, *J. chem. Phys.* **21,** 2056–2064 (1953).
61. F. J. Rohlf, Adaptive hierarchical clustering schemes, *Syst. Zool.* **19,** pp. 58–82 (1970).
62. F. J. Rohlf and D. R. Fisher, Tests for hierarchical structure in random data sets, *Syst. Zool.* **17,** 407–412 (1968).
63. A. K. Romney, R. N. Shepard and S. B. Nerlove, *Multidimensional Scaling*, Vol. 2 (applications). Seminar Press, New York (1972).
64. J. Rubin, Optimal classification into groups: an approach for solving the taxonomy problem, *J. Theor. Biol.* **15,** 103–144 (1967).
65. R. Saunders and G. M. Funk, Poisson limits for a clustering model of Strauss, *J. appl. Prob.* **14,** 776–784 (1977).
66. J. Schultz and L. Hubert, Data-analysis and connectivity of random graphs, *J. math. Psychol.* **10,** 421–428 (1973).
67. S. L. Sclove, Population mixture-models and clustering algorithms, *Communs. Statist. Theory and Meths.* **A6,**

417–434 (1977).

68. A. J. Scott and M. J. Symons, Clustering methods based on likelihood ratio criteria, *Biometrics* **27**, 387–397 (1971).
69. G. Sebestyen and J. Edie, An algorithm for non-parametric pattern recognition, *IEEE Trans. Elect. Comput.* **EC-15**, 908–915 (1966).
70. R. J. Shanley and M. A. Mahtab, Delineation and analysis of clusters in orientation data, *Math. Geol.* **8**, 9–23 (1976).
71. R. N. Shepard, A. K. Romney and S. B. Nerlove, *Multi-Dimensional Scaling*, Vol. 1 (Theory). Seminar Press, N.Y. (1972).
72. P. H. A. Sneath, Method for testing distinctness of clusters – Test of disjunction of 2 clusters in euclidean space as measured by their overlap, *Math. Geol.* **9**, 123–143 (1977).
73 P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*. W. H. Freeman, San Francisco (1973).
74. R. R. Sokal, Classification: purposes, principles, progress, prospects, *Science* **185**, 1115–1123 (1974).
75. D. J. Strauss, Model for clustering, *Biometrika* **62**, 467–475 (1975).
76. J. S. Strauss, J. J. Bartko and W. J. Carpenter, Jr., The use of clustering techniques for the classification of psychiatric patients, *Br. J. Psychiat.* **122**, 531–540 (1973).
77. J. S. Strauss, Classification by cluster analysis, *Report of the International Pilot Study of Schizophrenia*, Vol. 1, Chapter 12, pp. 336–359. World Health Organization, Geneva (1973).
78. A. I. Torn, Cluster analysis using seed points and density-determined hyperspheres as an aid to global optimization, *IEEE Trans. syst. man Cybernet.* **SMC-7**, 610–616 (1977).
79. O. I. Tosi, The problem of speaker identification and elimination, *Measurement Procedures in Speech, Hearing and Language*, S. Singh, ed., University Park Press, Baltimore (1975).
80. M. E. Turner, Credibility and cluster, *Ann. N.Y. Acad. Sci.* **161**, 680–688 (1969).
81. C. J. Van Rijsbergen, A clustering algorithm, *Comput J.* **13**, 113–115 (1970).
82. R. F. White and T. M. Lewinson, Probabilistic clustering for attributes of mixed type with biopharmaceutical applications, *J. Am. Statist. Ass.* **72**, 271–277 (1977).
83. M. Wirth, G. F. Estabrook and D. J. Rogers, A graph theory model for systematic biology with an example for the Oncidiinae (Orthidadeae), *Syst. Zool.* **15**, 59–69 (1966).
84. J. H. Wolfe, Pattern clustering by multivariate mixture analysis, *Multivar. behavior. Res.* **5**, 329–350 (1970).
85. W. E. Wright, A formalization of cluster analysis, *Pattern Recognition* **5**, 273–282 (1973).
86. L. A. Zadeh, Fuzzy sets, *Inf. Cont.* **8**, 338–353 (1965).
87. S. Zahl, A comparison of three methods for the analysis of spatial patterns, *Biometrics* **33**, 681–692 (1977).
88. C. T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Comput.* **C-20**, 68–86 (1971).
89. A. Cohen, R. Gnanadesikan, J. R. Kettenring and J. M. Landwehr, Methodological developments in some applications of clustering, *Applications of Statistics*, P. R. Krishnaiah, ed., p. 141–162. North Holland Publishing (1977).
90. R. Gnanadesikan, J. R. Kettenring and J. M. Landwehr, Interpreting and assessing the results of cluster analyses, *Bull. int. Statist. Inst.* **47**, 451–463 (1977).

About the Author – RICHARD C. DUBES was born in Chicago, Illinois in 1934. He received his B.S. degree from the University of Illinois in 1956, his M.S. degree from Michigan State University in 1959 and his Ph.D. degree from Michigan State University in 1962, all in electrical engineering. In 1956 and 1957, he was a member of the technical staff of the Hughes Aircraft Company, Culver City, California. From 1957–1968, he served as graduate assistant, research assistant, Assistant Professor and Associate Professor in the Electrical Engineering Department at Michigan State University. In 1969, he joined the Computer Science Department at Michigan State University and became Professor in 1970. He is the author of *The Theory of Applied Probability* (Prentice-Hall, 1968) and several technical papers and reports.

Dr. Dubes has served as a consultant to the Lear-Siegler Corp., Grand Rapids, Michigan and the J. M. Richards Laboratory, Detroit, Michigan. His areas of technical interest include pattern recognition, clustering, decision theory and application of data analysis methods to the medical area. He is a member of the Institute of Electrical and Electronic Engineers, the Pattern Recognition Society and Sigma Xi.

About the Author – ANIL K. JAIN was born in Basti, India on 5 August, 1948. He received his B. Tech. degree with distinction from Indian Institute of Technology, Kanpur in 1969 and his M.S. and Ph.D. degrees in electrical engineering from Ohio State University, Columbus, in 1970 and 1973, respectively. He was recipient of the National Merit Scholarship in India.

From 1971 to 1972, he was a Research Associate at the Communications and Control Systems Laboratory, Ohio State University. Then, from 1972 to 1974, he was an Assistant Professor in the Computer Science section at Wayne State University, Detroit. In 1974, he joined the Computer Science Department at Michigan State University where he is currently an Associate Professor. His research interests are in the areas of pattern recognition and image processing.

Dr. Jain is a member of the Association of Computing Machinery, the Institute of Electrical and Electronic Engineers, the Pattern Recognition Society and Sigma Xi.