

Chapter 9

CLUSTERING

KLAUS KRIPPENDORFF

The Annenberg School of Communications
University of Pennsylvania
Philadelphia, Pennsylvania

I. Clustering	259
II. Canonical Form of Data	261
III. Relevance of Data	262
IV. Ordinality of Data	263
V. Derived Form of Data: Distances	266
VI. Goals and Computational Efforts	272
VII. Presentation of Results	275
A. Dendrograms	275
B. Spatial Representations for Biordinal Clustering	276
C. Recorded Distance Matrices	278
D. Prototypes and Centroids	279
E. Other Multivariate Techniques	281
VIII. Properties of Emerging Clusters	281
IX. Clustering Algorithms	289
A. The Johnson Algorithm	289
B. The CONCOR Algorithm	291
C. The Strong Association Algorithm	292
D. The Multivariate Classification Algorithm	296
X. Validation and Vindication	300
XI. Summary—Conclusion	305
References	307

I. CLUSTERING

Clustering seeks to group or to lump together objects or variables that share some observed qualities or, alternatively, to partition or to divide a set of objects or variables into mutually exclusive classes whose boundaries reflect differences in the observed qualities of their members. Clustering thus extracts typologies from data which in turn represent a reduction of data complexity and may lead to conceptual simplifications.

Clustering should not be confused either with the analysis of the groupings made by subjects or with the assignment of objects to the categories in which they belong.

Clusters emerge from the interaction between the characteristics that are manifest in multivariate data and the assumptions that are built into the procedure. The recognition of these assumptions pertains to problems of validity, to which much of the paper and its conclusion are devoted.

Clustering originated in anthropology (Driver and Kroeber, 1932) and in psychology (Zubin, 1938; Tryon, 1939) in response to the need for empirically based typologies of cultures and of individuals. Computational problems hindered the initial development of these ideas. But by the early 1960s clustering techniques emerged in a variety of other disciplines, including biology (Sokal and Sneath, 1963). Applications are now so numerous that references to them would fill a book (see Sneath and Sokal, 1973). Problems to which clustering has found answers range from counting dust particles and bacteria, to land allocation in urban planning and political campaigning. Clustering has proven useful especially in psychology, anthropology, sociology, political science, economics, management, geography, and literature—virtually the whole spectrum of the behavioral and social sciences (Bailey, 1975) in which data do not exhibit the determinism of the natural sciences and theories are based on types, categories, and differentiations that knowingly omit some of the less significant variations in the observed phenomena.

In communication research, clustering provides a valuable tool for identifying cliques from sociometric or communication network type data, for example, or for detecting “invisible colleges” as manifest through citations of literature in scientific publications. Clustering is also used for grouping concepts that appear highly associated in given messages into stereotypes, for developing emic as opposed to etic type categories for content analysis from receiver responses, or for detecting redundant questionnaire items that may be explained by a common underlying variable. Clustering may also be used to simplify the representation of complex communication systems and thus provides the pretext for other forms of analysis including modeling.

As clustering has been applied to more and more diverse subject areas, clustering procedures have, themselves, grown in variety. I will not attempt to present a survey of either. Rather, based on the belief that all clustering techniques follow a few basic principles, with ample room open for further applications and developments of details, I shall discuss some of the options an analyst faces when deciding among existing clustering procedures or when assembling one for his special purpose, and I shall discuss some of the implications such choices have regarding computational efforts, validity and interpretability of results. This chapter provides in a sense a collection of *tools for evaluating what exists* and for *constructing anew what is needed* when multivariate data are to be analyzed by what has become known as clustering.

		variables					
		1	2	...	u	...	m
objects	1						
	2						
	⋮						
	i				x_{iu}		
	⋮						
	n						

FIG. 1. An $n \times m$ matrix of data in canonical form.

II. CANONICAL FORM OF DATA

Figure 1 depicts the canonical form of data for clustering. It is an $n \times m$ matrix X with entries x_{iu} denoting measures of some sort. The common interpretation of this matrix is that each of n objects is described in terms of m values, each pertaining to a different variable. The rows of this matrix are m -tuples or vectors with m components. The variables may have different metrics, nominal, ordinal, interval, and ratio metric, and may have any number of degrees of freedom. This includes binary variables as a special case.

It is basic to clustering that a matrix X whose variables have the *same metric throughout* can be interpreted in two ways, as objects \times variables and as a variables \times objects, for it is then possible to cluster objects in terms of variables, variables in terms of objects, and indeed both in terms of each other. For example, the matrix in Table I can be depicted either as in Fig. 2 or as in Fig. 3. Figure 2 depicts distances between objects as would be required when objects are clustered in terms of the values on their descriptive variables. In Fig. 3 distances between variables are depicted in an object space with variables X and Y shown to be in close proximity.

TABLE I

DATA MATRIX WITH RATIO METRIC ENTRIES

		variables		
		X	Y	Z
objects	A	2	2	0
	B	1	0	3
	C	1	2	3

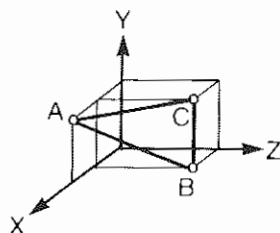


FIG. 2. Distances between objects in a space of variables.

An important distinction is whether data are "complete." If some of the entries x_{iu} are not known or are unavailable for analysis, data are incomplete and require that special assumptions be adopted to compensate for the missing entries. This chapter considers complete data only.

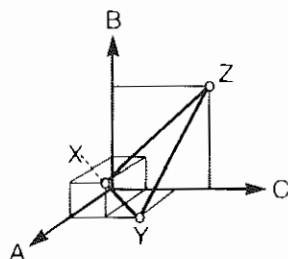


FIG. 3. Distances between variables in a space of objects.

III. RELEVANCE OF DATA

It is important that the variables that are chosen as descriptive of the objects in a sample be relevant to the attempted clustering. Individuals can be categorized in terms of their income, occupation, and social status, in terms of psychopathologies, in terms of physical conditions, their weight, height, strength, in terms of their life styles, etc. The choice among descriptive variables depends on the purpose of the clusters that are expected to emerge. While sampling theory provides statistical criteria for choosing among the *objects* of a population, criteria are less clear for choosing among the potentially infinite universe of possible *variables*.

Generally, variables that either vary randomly or remain constant in the data may be ignored in cluster analysis as they provide little help in differentiating among objects. Also redundant variables (different measures of the same underlying dimension) should be avoided for they only increase computational efforts. The identification of constant, random, and redundant variables can be accomplished by a variety of analytical techniques. For example, factor analysis has been used to identify orthogonal variables. These could be said to be least redundant.

Except for the removal of constant, random, and redundant variables, the researcher requires a theory or at least good intuition to decide on the relevance of the remaining variables to a given problem. For example, a researcher wishing to cluster psychotherapeutic patients for the purpose of standardizing clinical treatments should derive his variables from existing theories about such treatments. Similarly, a researcher attempting to develop typologies for social organization of industrial enterprises must be careful to include all of the variables that sensible writers have associated with industrial organizations. In using theoretical writings to identify relevant variables, however, the researcher may encounter three basic difficulties. First, different theories may be concerned with different levels of aggregation, for example, alienation, an individual's state, versus vertical organization, a characteristic of formal structure in which individuals take part. Second, theoretical concepts may be abstract and multidimensional. Authoritarianism, for example, may not be measurable by a single variable. Finally, variables may not have the same weight. For example, "education" has more influence than "sex" on the formation of social groups focusing on academic topics, but the reverse is true for humor, social stereotypes, and economic expectations.

In summary, the researcher should choose variables which are on the same level of abstraction, equal in weight, and logically independent of each other. But most of all such variables should feed into a theory or conceptual system that renders the description of the objects of analysis meaningful specimens for clustering.

Though a lack of relevance may greatly complicate the interpretation of clustering results, this is a problem that is extraneous to the process of clustering and can therefore be mentioned only in passing. Section X will examine a second source of difficulties of interpretation. The following Sections IV and V are concerned with properties and forms of data.

IV. ORDINALITY OF DATA

Like all other multivariate techniques, clustering methods are used in identifying certain patterns within available data and differ mainly in the way they define, recognize, and represent such patterns.

When clustering is defined as "a technique for grouping objects that are in close *proximity* to each other," one has a biordinal (of the order two) conception of the pattern in mind that clusters are to represent. Biordinal techniques either accept or immediately convert data into distances, differences, similarities, disagreements, correlations, etc. Distances, etc., are quantitative expressions for relations between *two* objects or between *two* variables. They have exactly *two* arguments and belong to the class of *binary* relations. Biordinal clustering procedures yield clusters whose members stand to each other in a certain *pairwise* relationship, proximity being one example.

TABLE II
DATA MATRIX WITH BINARY ENTRIES

		variables			
		1	2	3	4
objects	A	1	1	0	0
	B	1	0	1	0
	C	0	1	1	0

There exist more complex relations, however, that cannot be decomposed without loss into a set of binary relations. This is easily demonstrated and provides the ground for the distinction between *biordinal* and *multiordinal* clustering techniques.

Suppose one is given a 3 by 4 matrix X as in Table II with binary attributes, 0 or 1, in all cells.

The distances or associations between the pairs of objects in Table II will have to be defined on three two-dimensional contingency matrices depicted as Table III. The uniform distribution of probabilities in these matrices indicates that the pairwise co-occurrence of attributes may be due entirely to chance. A biordinal clustering technique would therefore find no justification for merging objects into clusters. However, when objects are examined in triples rather than in pairs, one finds a strong tertiary relation present. This becomes obvious when the data in Table II are represented three-dimensionally as in Table IV.

An example of a relation between three objects that is fully explainable in terms of any two of its three component binary relations is given as Table V. Here, biordinal clustering would be perfectly justifiable for there is nothing unique about the combination of the three objects that could not be expressed in binary terms.


TABLE III
CONTINGENCY MATRICES CONTAINING NO ASSOCIATION IN PAIRS

		B		C		C	
		0	1	0	1	0	1
A	0	.25	.25	A	0	.25	.25
	1	.25	.25		1	.25	.25
B	0	.25	.25	B	0	.25	.25
	1	.25	.25		1	.25	.25

TABLE IV
CONTINGENCY MATRIX CONTAINING A TERTIARY ASSOCIATION

TABLE V

CONTINGENCY MATRICES CONTAINING PAIR ASSOCIATIONS ONLY



A 3D cube with axes labeled A (vertical), B (horizontal to the left), and C (depth to the right). The value .5 is written at each of the eight corners of the cube.

B

.5	0
0	.5

A

N

C

.5	0
0	.5

A

C

.5	0
0	.5

B

V. DERIVED FORM OF DATA: DISTANCES

Biordinal clustering usually starts with data in the form of a square matrix D whose entries d_{ij} measure some distance, difference, or dissimilarity either (and generally) between all pairs of objects or (provided they possess the same metric) between all pairs of variables. See Fig. 4.

Distances between the same objects must be zero. Otherwise, distances must be positive and symmetrical:

$$d_{ii} = 0, \quad d_{ij} \geq d_{ii}, \quad d_{ij} = d_{ji}$$

In order to possess at least interval metric properties within many-dimensional space, distances must also satisfy the triangle inequality:

$$d_{ik} \leq d_{ij} + d_{jk}$$

and in some cases the ultrametric inequality (Jardine and Sibson, 1971; Johnson, 1967):

$$d_{ik} \leq \max(d_{ij}, d_{jk})$$

Data may also be represented through measures of similarity, agreement, resemblance, or correlation, s_{ij} . Similarities and distances are inversely related with the least similar objects giving rise to large distances and small distances reflecting strong resemblances. Similarity measures may be converted into distance measures, for example by

$$d_{ij} = s_{ii} - s_{\min} - s_{ij}$$

$$d_{ij} = (s_{ii} - s_{\min} - s_{ij})^{1/2}$$

$$d_{ij} = (s_{ii} - s_{ij}) / (s_{ii} - s_{\min})$$

$$d_{ij} = [(s_{ii} - s_{\min})(s_{ii} - s_{ij})]^{1/2}$$

	objects					
	1	2	...	j	...	n
1	d_{11}	d_{12}	...	d_{1j}	...	d_{1n}
i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{in}
:	:	:		:		:
:	:	:		:		:
n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{nn}

FIG. 4. A n^2 matrix of distances.

The latter is my generalization of Gower's (1966) transformation, developed to convert similarity measures that range between 0 and +1. A useful conversion for correlation coefficients has been suggested by Tukey (1977):

$$d_{ij} = 1 - r_{ij}^2$$

It expresses the degree to which two objects are linearly related (positive or negative) and takes its maximum value when the objects are statistically independent. This serves as an example that distances and similarities can have many different interpretations which need to be understood before a clustering is attempted. Motivations for these conversion formulas cannot be given here.

I shall now give several distance and similarity measures and show how some of the more familiar measures can be regarded as special cases. For this purpose I shall first define four kinds of differences, one for each metric, then present methods of standardizing such differences across different metrics, and finally present a few key distance and similarity measures.

Difference notions depend on the metric of the variables involved. This may be seen in the comparison between nominal and interval data. Nominal data are characterized by qualitative distinctions without any implied order. Thus a nominal value matches with another or it does not and all mismatching pairs differ to the same degree. Interval data, on the other hand, recognize an ordering of values that allows additions and subtractions. Differences then become a function of their algebraic difference and may be large or small. These intuitive notions can be given rigorous forms: For *nominal* scales the difference between two values x_{iu} and x_{ju} of the u th variable is

$$\Delta_{ij, u} = \begin{cases} 0 & \text{iff } x_{iu} = x_{ju}, \\ 1 & \text{iff } x_{iu} \neq x_{ju} \end{cases}$$

For *ordinal* variables in which merely the rank orders count, such a difference is a function of the number of ranks above and below the two values to be compared. With

$$x_{iu}^* = 1 + \frac{1}{n} \left[\sum_{x_{ku} > x_{iu}} n_{x_{ku}} - \sum_{x_{ku} < x_{iu}} n_{x_{ku}} \right]$$

and x_{ju}^* defined analogously, the difference in ordinal scales becomes

$$\Delta_{ij, u} = |x_{iu}^* - x_{ju}^*|$$

For variables with *interval* metric the difference is as just discussed:

$$\Delta_{ij, u} = |x_{iu} - x_{ju}|$$

and for variables with *ratio* metric, the difference may be expressed by

$$\Delta_{ij, u} = |x_{iu} - x_{ju}| / (|x_{iu}| + |x_{ju}|)$$

If both x_{iu} and x_{ju} are positive, as should be expected in ratio-level measurements, then the ratio difference is a modification of Lance and Williams'

(1967) Canberrametric. All of these differences are taken from Krippendorff (1973), where also the motivations for their form may be found.

The most obvious way of aggregating differences across variables into a measure of *distance* is by summing some power r of it:

$$d_{ij} = \left[\sum_{u=1}^m (\Delta_{ij,u})^r \right]^{1/r}$$

When $r = 1$, differences $\Delta_{ij,u}$ are assigned equal weights and are merely summed. When variables are moreover dichotomous, d_{ij} becomes the Hamming distance (Hamming, 1950). For $r = 2$, d_{ij} corresponds to the familiar Euclidean distance in multidimensional space which has been used since Heinecke (1898). The distance is common in research on the semantic differential (Osgood, Suci, and Tannenbaum, 1957) and has been discussed in the cluster analysis literature by Sokal and Sneath (1963), Gower and Ross (1969), and many others. Generally an increase in the exponent r increases the impact of larger differences over smaller ones and thereby affects the nature of the clusters formed.

The Euclidean distance is appropriate when all variables possess the same metric but not when the metric of the variables differs. Variables then will have to be standardized. There seem to be three *methods of standardizing distances* across different metrics.

The *first* is a *reduction* of the power of the metric of all variables to the *least powerful metric* among the variables. The possible metrics may be listed in order of increasing power: nominal, ordinal, interval, ratio. Thus if there are ratio scales (e.g., numerical age, income in dollars) and ordinal scales (e.g., variables containing such values as "most conservative," "somewhat conservative," "neutral"), the values of all variables would then have to be regarded merely by their rank within the set of all values. Similarly, when binary attributes occur, all variables would then have to be dichotomized (e.g., above or below a certain age, income, liberal versus conservative).

A *second* method is to transform all values x_{iu} so that their *range* falls within the interval 0 and 1 (Sokal and Sneath, 1963; Gower, 1971). The transformed value may be expressed as

$$x'_{iu} = \frac{x_{iu} - \min_u(x_{ku})}{\max_u(x_{ku}) - \min_u(x_{ku})}, \quad k = 1, \dots, n$$

where $\min_u(x_{ku})$ and $\max_u(x_{ku})$, $k = 1, \dots, n$, represent the smallest and largest values in u , respectively. This method, however, is inapplicable to nominal metric variables, and if applied to ordinal data, it would assume interval characteristics that are not there.

A *third* method, and one that I prefer, is to standardize the *variance* in each variable before summing. For this purpose one computes a variable-specific

weight:

$$w_{u(r)} = \left[\sum_{i=1}^n \sum_{j=1}^n (\Delta_{ij, u})^r \right]^{-1}$$

so that

$$\hat{\sigma}_u^r = \sum_{i=1}^n \sum_{j=1}^n w_{u(r)} (\Delta_{ij, u})^r = 1$$

for each variable. The weighted distance then follows the form proposed by Bock (1974):

$$d_{ij} = \left[\sum_{u=1}^m w_{u(r)} (\Delta_{ij, u})^r \right]^{1/r}$$

Evidently, when $r = 2$ and all $w_{u(2)} = 1$, the distance becomes again the Euclidean distance.

Another generalization of the Euclidean distance has been proposed by Mahalanobis (1936). For $n > m$, let the $m \times m$ covariance matrix Σ have entries:

$$\sigma_{uv} = \frac{1}{n} \sum_{i=1}^n (x_{iu} - \bar{x}_{.u})(x_{iv} - \bar{x}_{.v})$$

where $\bar{x}_{.u}$ and $\bar{x}_{.v}$ are the means in variable u and v , respectively. Let the inverse of this matrix Σ^{-1} have entries σ^{uv} , then the Mahalanobis distance is defined by

$$d_{ij} = \left(\sum_{u=1}^m \sum_{v=1}^m \sigma^{uv} \Delta_{ij, u} \Delta_{ij, v} \right)^{1/2}$$

It is noticeable that the Mahalanobis distance depends on all n objects simultaneously with its range in fact a function of n :

$$0 \leq d_{ij} \leq (2n)^{1/2} \quad \text{and} \quad \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 = 2mn^2$$

The Mahalanobis distance also eliminates the effects of possible correlations among pairs of variables on the distance between two objects. And it is relatively independent of the ranges in each variable. The Mahalanobis distance is not applicable across different metrics, but this drawback might be corrected by standardizing the variance in each variable according to the third method just discussed. Accordingly, the corrected covariance matrix Σ^* has then entries

$$\sigma_{uv}^* = w_{u(1)} w_{v(1)} \sigma_{uv}$$

and the corrected distance becomes

$$d_{ij} = \left(\sum_{u=1}^m \sum_{v=1}^m \sigma^{uv} w_{u(1)} w_{v(1)} \Delta_{ij, u} \Delta_{ij, v} \right)^{1/2}$$

A most common similarity measure is the product-moment correlation coefficient r_{ij} which is defined by

$$r_{ij} = \frac{\sum_u x_{iu} x_{ju} - \bar{x}_i \bar{x}_j}{(\sum_u x_{iu}^2 - \bar{x}_i^2)(\sum_u x_{ju}^2 - \bar{x}_j^2)}$$

where \bar{x}_i and \bar{x}_j are the mean values for objects i and j , respectively. This reliance on mean values presupposes that all variables possess interval or ratio metrics.

While the correlation coefficient can easily be converted into a distance measure by one of the conversion formulas, its interpretation as a similarity measure is not too clear. $r_{ij} = 0$ denotes statistical independence, and $|r_{ij}| = 1$ denotes that variables are linearly dependent. Nonlinear relationships and relations of higher ordinality reduce the value of $|r_{ij}|$, however. The difference between positive and negative values of r_{ij} adds another difficulty to the interpretation of the coefficient as a similarity measure. For example, it is not too obvious whether two objects between which a strong negative linear relation exists should thereby be regarded as similar or different.

Opinions on the use of correlation coefficients for clustering vary considerably in the literature. (For additional arguments see Section X.)

In an approach to multiordinal clustering of nominal data, Krippendorff (1969, 1974) used information theoretical measures to assess the loss of structure in many-dimensional spaces caused by the grouping of the objects' qualitative descriptions. The notion of structure here considered may be paraphrased by "interdependent differentiation," "trans-information," "multiple-order interaction," or "relational entropy" and might be said to be the opposite of redundancy and randomness. The loss in the amount of structure due to the elimination of qualitative distinctions in one or more variables can be expressed by several distances between two objects. I shall define only one here. Let $n_{\langle i \rangle}$ be the number of objects with the description $\langle x_{i1}, \dots, x_{iu}, \dots, x_{im} \rangle$, n_{x_u} be the number of objects described in terms of x in the u th variable, and $n_{x_{iu}}$ be the number of objects that share the value x in the u th variable with the object i . The total amount of structure within the m -dimensional space—that is, the total amount of relatedness manifest in the m -valued distribution of objects—is

$$T = \sum_i \frac{n_{\langle i \rangle}}{n} \log_2 \frac{n_{\langle i \rangle}}{n} - \sum_{u=1}^m \sum_x \frac{n_{x_u}}{n} \log_2 \frac{n_{x_u}}{n}$$

And the distance between two objects i and j becomes the loss in structure when all hyperplanes within which i and j are located are to be merged.

To express this, the two m -tuples $\langle x_{i1}, \dots, x_{iu}, \dots, x_{im} \rangle$ and $\langle x_{j1}, \dots, x_{ju}, \dots, x_{jm} \rangle$ are considered to be composed of two parts i_C and $k_{\bar{C}}$ and j_C and $k_{\bar{C}}$, respectively, where C is a set of variables within which $\langle i \rangle$ and $\langle j \rangle$ differ. With a loss function defined by

$$\text{Loss}(a, b) = \begin{cases} 0 & \text{iff } a = b \text{ or } n_a = 0 \text{ or } n_b = 0, \\ (n_a + n_b) \log_2(n_a + n_b) - n_a \log_2 n_a - n_b \log_2 n_b & \text{otherwise} \end{cases}$$

the loss in structure, expressed as a distance, becomes

$$d_{ij} = T(\text{before}) - T(\text{after merging } i \text{ and } j) \\ = \frac{1}{n} \left[\sum_C \sum_{k_{\bar{C}}} \text{Loss}(\langle i_C k_{\bar{C}} \rangle, \langle j_C k_{\bar{C}} \rangle) - \sum_{u=1}^m \text{Loss}(x_{iu}, x_{ju}) \right]$$

This distance is small when objects are redundant, i.e., have many values in common, and values with respect to which the objects differ carry little information. The distance is large when the descriptions of objects represent significant differentiations within the m -dimensional space. Like the Mahalanobis distance, the preceding takes all objects into consideration that share some values with either of the two objects being compared. But, unlike the Mahalanobis distance, it takes account of the multiordinal nature of the distribution of objects and does not assume any ordering of values.

For binary attributes, 0 or 1, in all variables several simple distances have been used. To simplify the notation, let me represent the matching and mismatching of attributes associated with objects i and j in terms of a 2 by 2 contingency table:

		j		
		1	0	
i	1	a	b	e_i
	0	c	d	$1 - e_i$
		e_j	$1 - e_j$	1

$$a = \frac{1}{m} \sum_{u=1}^m x_{iu} x_{ju}, \quad e_i = \frac{1}{m} \sum_{u=1}^m x_{iu}$$

where a is the proportion of matching ones between i and j , b is the proportion of mismatches with i s zeros co-occurring with j s ones, etc. (the distance d_{ij} is not to be confused with the proportion d). In these terms, the Euclidean distance becomes

$$d_{ij} = (b + c)^{1/2}$$

and two simple matching coefficients used by Zubin (1938) and Jaccard

(cited in Sokal & Sneath, 1963), respectively, are

$$d_{ij} = b + c \quad \text{and} \quad d_{ij} = a + b + c$$

Yule's association coefficient

$$s_{ij} = (ad - bc)/(ad + bc).$$

has also been used as a similarity measure (Sokal & Sneath, 1963).

Space does not permit a review and discussion of the many distance and similarity measures which are possible and have actually been applied in clustering. The user of any clustering technique that starts from distance or similarity data must ascertain though that these measures possess the metric properties that the clustering technique requires and that the assumptions implied by the choice of a particular distance measure conform to what the clusters are expected to represent. Specifically, multiordinal relations are not manifest in distance or similarity data. Multiordinal clustering techniques require continuous interaction with data in their canonical form because they are capable of retaining such relations.

VI. GOALS AND COMPUTATIONAL EFFORTS

A major problem of all multivariate techniques is the amount of computation required to produce results. Since the number of cells in a many-dimensional space grows exponentially with the number of dimensions, such numbers often approach limits of computability before data can be considered rich enough to contain interesting information. Virtually all multivariate analysis algorithms rely on computational shortcuts to reduce this effort and thereby impose assumptions on the way data are processed. Cluster analysis is no exception and the user should know what is involved.

While all clustering procedures yield groupings of objects or variables according to some criterion, the specific task may be one of the following:

(a) selecting that subset of a set of objects which contains a designated object in relation to which criteria for inclusion into the subset, class, type, or cluster are defined.

(b) selecting that partition of a set of objects into a specified number of exhaustive and mutually exclusive subsets, classes, types or clusters, the parts of the partition, of which each is in a specifiable sense optimal under the numerical restriction.

(c) selecting that partition of a set of objects into any number of exhaustive and mutually exclusive subsets which satisfies a specified criteria of optimality.

(d) selecting that binary decision tree which contains only partitions satisfying (b) including the partition satisfying (c).

Table VI shows how the number of alternatives among which decisions need to be computed grows with the number of objects to be clustered. With only $n = 10$ objects, task (a) presents 512 alternatives, which is a manageable and unproblematic number. But task (c) requires the evaluation of 116 thousand partitions, a number that approaches practical limits of computation, while task (d) with its 2.6 billion decision trees already exceeds computational limits. The lesson to be drawn from any extension of numbers n of this table into the domain of theoretical significance is that practical clustering procedures cannot compute alternatives *simultaneously* but must proceed *iteratively*, in irreversible steps. It is in the *form* of iteration that *hierarchical clustering schemes* emerge.

Two iterative clustering procedures can be distinguished: The successive *partitioning* of a set of objects into more and smaller subsets (classes or clusters) and the successive *merging* of objects into fewer and larger subsets (classes or clusters). Sneath and Sokal (1973) call the former technique *divisive* and the latter *agglomerative*. With a vertical bar separating the parts of a partition among four objects, a , b , c , and d , these two options are depicted in Fig. 5 as a path through a partition lattice from top to bottom or from bottom to top, respectively.

While either option results in the choice of one out of $n!(n-1)!2^{n-1}$ binary decision trees, their computational efforts are rather different as the following table of the number of alternatives may show:

	1st step	2nd step	...
Successive partitioning:	$2^{n-1} - 1$	between $2^{\frac{1}{2}(n-1)} - 1$ and $2^{n-2} - 1$...
Successive merging:	$\frac{n(n-1)}{2}$	$\frac{(n-1)(n-2)}{2}$...

Numerically, when $n = 100$, the first step of successive partitioning poses about 10^{30} alternative partitions to choose from, a practical impossibility, whereas successive merging calls for the evaluation of only 4950. While successive partitioning cuts these numbers down rather quickly, the first step is evidently prohibitive.

Finally, computation is affected by two further distinctions. The first was suggested by Lance and Williams (1967). They define a *combinatorial strategy* as one in which the original input matrix (of distances or of data in canonical form) is successively transformed, becomes smaller and simpler, and thereby reduces the computational effort by each step. In contrast, in a *noncombinatorial strategy* all computations are based on the original input which must therefore be maintained throughout. Obviously, combinatorial strategies are more efficient. Furthermore, I should like to distinguish between two combinatorial strategies: *distance-recursive* strategies in which each distance matrix is derived from its preceding distance matrix, and *data-recursive* strategies in

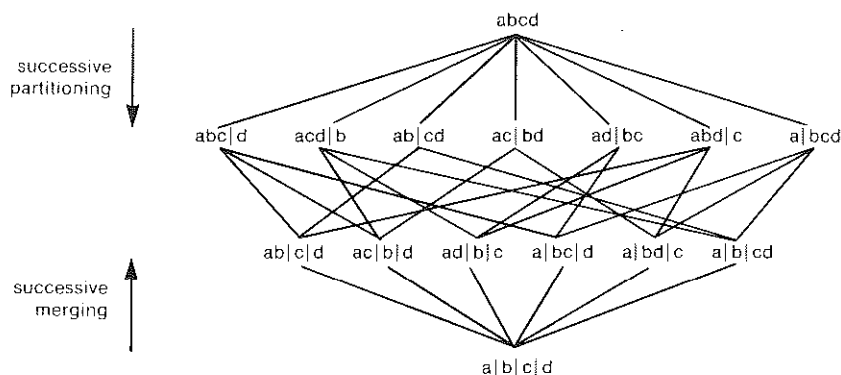


FIG. 5. Partition lattice involving four objects.

which it is the canonical representation of data which is successively modified. The latter is exemplified in Table VII. Distance recursive strategies are at least m times (m = the number of variables) as efficient as data recursive methods.

Of course, computational advantages must be weighted against the amount of information that combinatorial strategies lose. Computational shortcuts can not bypass questions of validity.

VII. PRESENTATION OF RESULTS

The conceptualization of multivariate data is difficult. We are just not accustomed to seeing point distributions in four-or-more-dimensional spatial representations and it is this fact that often serves as a motivation for applying multivariate statistical analyses. All multivariate techniques transform such data. Even though analytical results may appear simple, it is often difficult to relate the transforms of these data to the original observations. This section does not deal with questions of meaningfulness and of the adequacy of the transformations for producing a result. It is rather concerned with several ways the results of a clustering process might be visualized leaving the conceptualization of the process for Section IX.

A. Dendrograms

The most important form of representing clustering results is the dendrogram (see Fig. 6) which is a tree-like structure whose branches terminate at the objects being clustered. The lengths of its branches indicate differences in homogeneity, or heterogeneity within clusters being merged or partitioned. Dendrograms are nothing but a more sophisticated form of listing objects by their membership in clusters: Each horizontal cut through a dendrogram

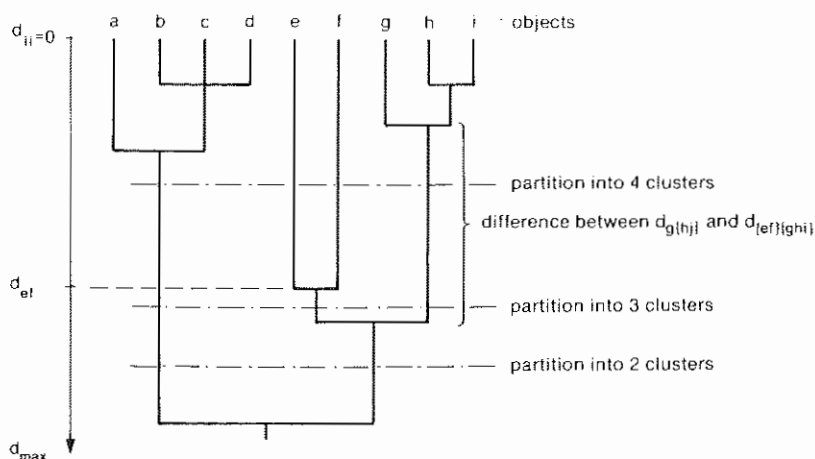


FIG. 6. Dendrogram representation of clustering.

indicates one of several partitions of the set of objects, the height between this cut and the original (unclustered) objects indicates the level of homogeneity or heterogeneity lost in the partition.

Dendrograms are particularly suited to represent a large number of objects and the whole history of a clustering process whether it proceeds by successive merging or partitioning. Relatively long stems between branching points indicate to the researcher where relatively large jumps in heterogeneity occur and might suggest cutoff points at which partitions might be meaningful.

Figure 7 represents a small section of a dendrogram obtained by clustering 299 sales appeals in television commercials (Dziurzynski, 1978) by the strong association method (see Section IX).

Johnson (1967) used a modified dendrogram which is particularly suited for computer printouts (Fig. 16).

B. Spatial Representations for Biordinal Clustering

By far the most appealing form of representation depicts the proximities among objects in some space and indicates clusters by drawing their boundaries. Since proximities are an essential ingredient of Gestalt perception, groupings are much easier to visualize when similarities, correlation, and the like are expressed as distances. Figures 8 and 9 exemplify such a representation in one and in two dimensions.

When three dimensions are involved, the representation is somewhat more cumbersome although still possible. Spatial representations in four or more dimensions become virtually unreadable however. Since many multivariate data consist of objects that are characterized by many more than three variables, the use of spatial representations of clustering results is extremely limited. But since two- or three-dimensional representations are so common,

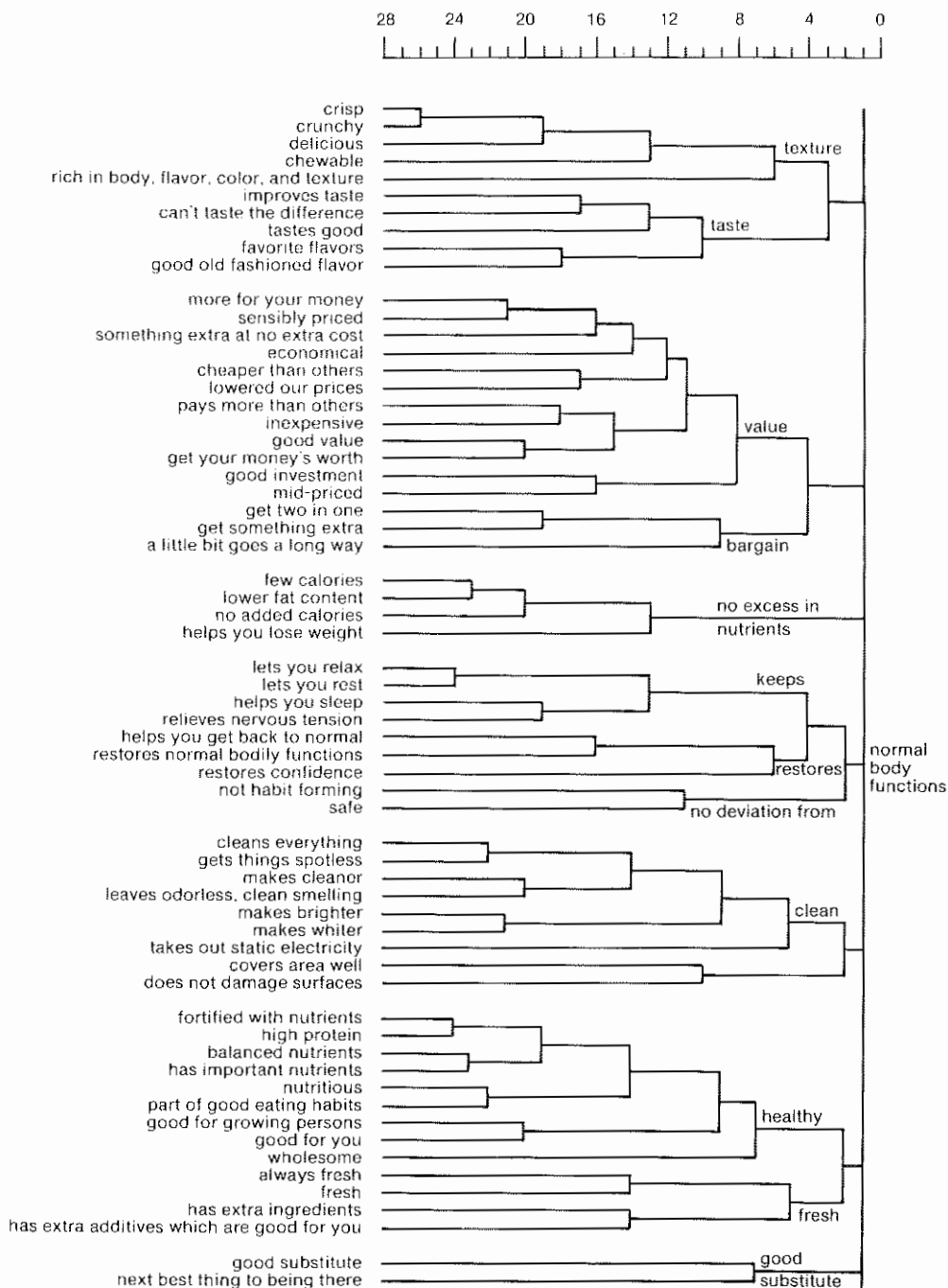


FIG. 7. Section of a dendrogram for television advertising appeals (after Dziurzynski, 1978).



FIG. 8. One-dimensional spatial representation of clusters.

researchers have either ignored the many-dimensional character of the distances between objects and approximated them by distances in two- or three-dimensional representations or else have employed dimension reducing techniques, such as factor analysis, that yield visually representable distributions of objects in a space with orthogonal dimensions. Since the size of the resulting clusters and distances between them are then distorted in the visual representation, numerical values for these may have to be added to indicate their true quantitative relationship. These are omitted in Fig. 10, which is the visualization of a taxonomy of the *Enterobacteriaceae* (Lysenko & Sneath, 1959).

It should be reemphasized that spatial representations with their emphasis on proximity carry strong biordinal biases. Higher-order relations among three or more objects have no obvious spatial form.

C. Reordered Distance Matrices

Several authors, among them Sneath and Sokal (1973), suggest that the results of distance recursive merging be represented by reordering the entries of the initial distance matrix D^0 so that the proximity of rows (and columns) reflect the rank ordering of distances between objects. However, when the initial distances are rearranged not by their rank but by the hierarchical ordering that any iterative clustering process imposes, one obtains a reordered distance matrix, such as in Fig. 11, in which entries are "blocked" into sections representing distances within and between clusters, respectively. The

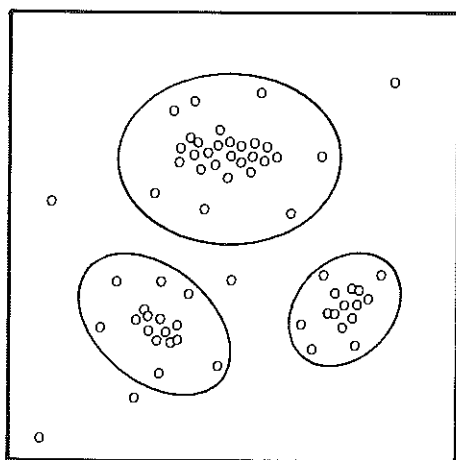


FIG. 9. Two-dimensional spatial representation of clusters.

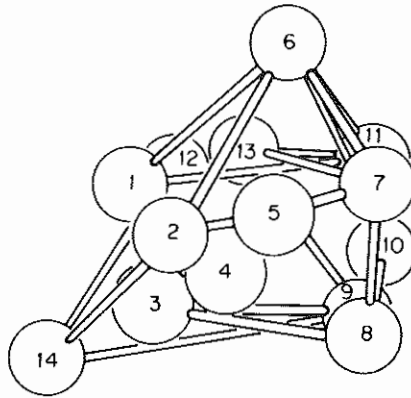


FIG. 10. Three-dimensional representation of many-dimensional clusters (Sneath & Sokal, 1963, with permission of W. H. Freeman and Company).

reordering of distance matrices does not reveal, however, the motivation for multiordinal clusters. It is instructive primarily when clustering proceeds from distances and is biordinal.

D. Prototypes and Centroids

It is often desirable to identify an object, real or hypothetical, that is most representative of a cluster. Such an object is called the prototypical object or centroid, respectively. *Centroids* locate a given cluster, in multivariate space. In monothetic clustering schemes, the centroid is that m -tuple of values that

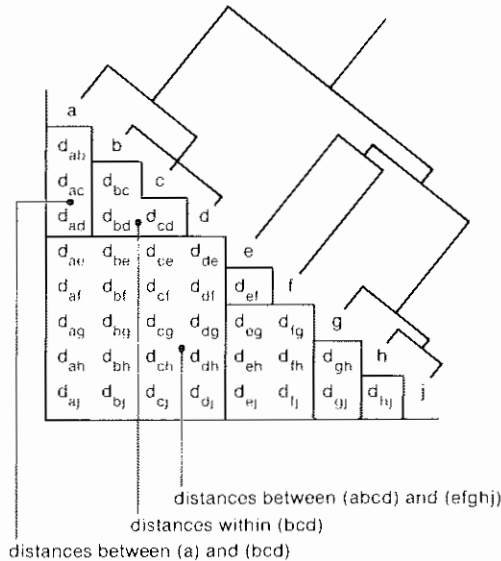


FIG. 11. "Blocked" distance matrix and associated dendrogram.

all objects in a cluster share. Since the objects in clusters emerging from polythetic techniques (see Section VIII) need not have any value in common, the centroid may then be quite abstract and is often unidentifiable in terms of data in canonical form. In either case, there may not exist a real object in the sample that coincides with the hypothetical centroid of a cluster. The object closest to the centroid of a cluster may then be chosen as the *prototype* of the cluster.

In single linkage clustering methods (see Section VIII) and those that employ measures of differences *between* rather than *within* clusters, centroids have no clear theoretical justification because clusters have this chain-like appearance and heterogeneities are not accounted for by these methods. The computation of centroids and the identification of prototypes is thereby not excluded however.

When data are in canonical form and biordinal clustering proceeds from distances, the centroid of the cluster G becomes the m -tuple

$$\langle \bar{x}_{G1}, \dots, \bar{x}_{Gu}, \dots, \bar{x}_{Gm} \rangle$$

where \bar{x}_{Gu} is the arithmetic mean of values of the u th variable over objects in cluster G when that variable has interval metric, the median when that variable has ordinal metric, and the mode when that variable is unordered.

Otherwise, the researcher tends to be restricted to identify the prototype of a cluster by

$$\min_{g \in G} (d_{gG}) \leftrightarrow g \text{ is the prototypical object of } G$$

in which d_{gG} stands for the heterogeneity measure chosen. In diametric clustering

$$d_{gG} = \max_{j \in G} (d_{gj})$$

whereby the prototype g is the one closest to the center of the circle circumscribing the objects in G . In variance-type clustering,

$$d_{gG} = \left(\frac{1}{n_G - 1} \sum_{j \in G} d_{gj}^r \right)^{1/r}$$

and when $r = 1$, the prototype occupies a position close to the mean of the cluster. Large values for r make that position more responsive to the skewness of the distribution of objects within a cluster.

In multivariate classification the prototype of a cluster is defined as that object which, when removed from the cluster, causes the least amount of structure loss within the cluster. With $n_{\bar{g}u} = n_{Gu} - n_{gu}$ and $n_{\langle \bar{g}_C k_{\bar{C}} \rangle} = n_{\langle G_C k_{\bar{C}} \rangle} - n_{\langle g_C k_{\bar{C}} \rangle}$ for simplicity of notation,

$$d_{gG} = \frac{1}{n} \left[\sum_C \sum_{k_{\bar{C}}} \text{Loss}(\langle \bar{g}_C k_{\bar{C}} \rangle, \langle g_C k_{\bar{C}} \rangle) - \sum_{u=1}^m \text{Loss}(\bar{g}_u, g_u) \right]$$

If the proportion $n_{\langle g_c k \bar{c} \rangle} / n_{\langle G_c k \bar{c} \rangle} = \text{constant}$ for all objects that share some values with g , then $d_{gG} = 0$ and g is both prototype and centroid of the cluster G .

Two drawbacks of representing clusters by prototypes are that there may exist objects that are actually more representative (closer to the centroid) of a cluster than those occurring in the sample and that there may exist many different objects for which d_{gG} is equal and minimum. The first problem is one of sampling and the second, one of measurement.

E. Other Multivariate Techniques

Insofar as it provides a single partition of canonically represented objects, clustering adds to the m descriptive values each object's membership in some cluster. Such an indication of membership can be regarded as the $(m + 1)$ st descriptive variable and thus expands the initial $n \times m$ matrix of data in canonical form to an $n \times (m + 1)$ matrix. This expanded matrix may be subjected to a variety of other multivariate techniques, for example, multiple discriminant techniques yielding explanations of the clusters in terms of the features that discriminate.

The results of clustering can be subjected to factor analysis to obtain a more efficient system of coordinates for representing these clusters (see Chapter 8). The results of different clustering techniques can be compared by cross-tabulations, etc. In fact, there is no limit to the use of other analytical techniques for describing and exploring the nature of the clusters that have been obtained as well as in preparing the data for subsequent clustering.

VIII. PROPERTIES OF EMERGING CLUSTERS

The aggregate or shared properties of objects within a cluster, the boundaries around clusters, the relations between clusters, the tree-like dendrograms describing either the history of merging objects into classes or the history of partitioning sets of objects into subsets are all expected to be based on given data. Once the criteria for iterative merging or partitioning are set, the clusters that do emerge develop certain properties that should not be an artifact of the procedure. Decision criteria for clustering must therefore be based on measures that characterize *sets of objects*. This section presents several measures on clusters of two or more objects and in terms of three dimensions of classification:

Difference measures versus heterogeneity measures

Single linkage measures versus multiple linkage measures

Polythetic measures versus monothetic measures

The distinction between biordinal and multiordinal clustering is also reflected in these measures and the formal conditions imposed on such measures depend to a large part on the computational approach taken, i.e., whether the procedure is distance recursive or data recursive and whether clusters are formed by partitioning (divisive) or by merging (agglomerative). Naturally the number of combinatorially possible clustering criteria exceeds those actually realized, and of those available only a few can be discussed here.

The differentiation between *difference* measures and *heterogeneity* measures is conceptually simple but the implications are far from obvious. Applied to clusters (classes of objects), difference measures quantitatively assess differences *between two* different clusters whereas heterogeneity measures assess differences *within one* cluster. Both are divergent generalizations of the distance between two objects. When a third object is added to a cluster of two, a difference measure assesses the difference between the third object and either one or both of the two objects already merged while a heterogeneity measure assesses some difference among all three objects regardless of how they were brought together and thereby assigns equal weight to each object involved.

Notationally, d_{EF} will be used to denote the *heterogeneity* of the union of two sets of objects, whereas $d_{E|F}$ will be used to denote the *difference* between the two sets. The distance d_{ij} between two objects i and j then is the special and overlapping case at which $E = \{i\}$ and $F = \{j\}$.

In these terms, and without reference to details, several formal requirements on the use of these measures as decision criteria for clustering can be stated. For *distance recursive* procedures, both heterogeneity and difference measures must satisfy analogous conditions:

$$\begin{array}{lll}
 d_{EE} \geq 0 & & \\
 d_{EF} \geq d_{EE}, & d_{E|F} \geq 0 & \text{Positive} \\
 d_{EF} = d_{FE}, & d_{E|F} = d_{F|E} & \text{Symmetrical} \\
 d_{EF} \leq d_{EG} + d_{GF}, & d_{E|F} \leq d_{E|G} + d_{G|F} & \text{Triangle inequality}
 \end{array}$$

to which the following may have to be added:

$$d_{EF} \leq \max(d_{EG}, d_{GF}), \quad d_{E|F} \leq \max(d_{E|G}, d_{G|F}) \quad \text{Ultrametric inequality}$$

These conditions correspond to those for distances between two objects except that the heterogeneity within a cluster, d_{EE} , may exceed zero and the difference between a cluster and itself, $d_{E|E}$, is meaningless by definition.

For *data recursive* procedures these formal requirements may be relaxed to the following two conditions:

(1) Decision criteria must not decrease with each upward move in the partition lattice (merging) and must not increase with each downward move in the partition lattice (partitioning). For any three clusters E , F , and G , at

any stage in the clustering process,

$$\max(d_{GG}) \leq \min(d_{EF})$$

For difference measures this does not apply generally. However, if one interprets $d_{G|G}$ as that distance which served as criterion to merge two clusters into G , then, while some differences within G may exceed $d_{E|F}$, at least under a successive merging strategy,

$$\max(d_{G|G}) \leq \min(d_{E|F})$$

This condition favors heterogeneity measures because whenever G is the union of E and F , $d_{GG} = d_{EF}$. For difference measures the condition must be rephrased to read: The largest of the differences that lead to the formation of a cluster must not exceed the smallest difference between any pair of clusters *at that stage at which that cluster was formed*.

(2) Ideally, decision criteria should also be independent of the order of cluster formation:

$$d_{\{ijk\}\{ijk\}} = d_{\{i\}\{jk\}} = d_{\{ij\}\{k\}} = d_{\{ik\}\{j\}}$$

This is again satisfied by heterogeneity measures but not by difference measures. Some of the implications of this failure will become apparent in the following.

The second dimension of classification refers to whether clusters are characterized by a *single representative linkage* between two objects or whether the measure aggregates *multiple linkages into a single index*. I shall exemplify the two dimensions by several individual and polythetic measures for biordinal clusters.

	Difference	Heterogeneity
Single linkage:	Connectedness	Diameter
Multiple linkage:	Average linkage	Variance

Clusters that are characterized by the *connectedness* of their members (Johnson, 1967; single linkage clustering according to Sokal & Sneath, 1963) stem from the simplest form of clustering with difference measures: Clusters are formed by merging objects in the increasing order of their distances, i.e., by merging those clusters that contain at least one object each, which is least distant *across* cluster boundaries. Thus, the two clusters E and F , if they are to be merged to form the cluster G , satisfy the criterion:

$$\min_{E, F} (d_{E|F}) = \min_{E, F} \left(\min_{i \in E, j \in F} (d_{ij}) \right)$$

The difference $d_{E|F}$ in this criterion is crucially dependent on the history of the formation of the cluster. It is the distance between two objects within G that at the point of the last addition to the cluster was the smallest distance

between objects of all different clusters formed prior to G . Within a cluster it is then possible to find distances between members that are larger or smaller than the distance on account of which the cluster was formed.

While leading to an extremely simple form of computation (in fact it can be done manually by inspection of the distance matrix D without any arithmetical computation), the weakness of this criterion lies in its tendency to form long chains that often bridge otherwise perfectly meaningful clusters. Some of the peculiar clusters that this criterion will produce are illustrated in the spatial representation of Fig. 12. The use of difference measures generally and in conjunction with single linkage conceptions of those measures in particular allows clusters to "grow out of control."

Clusters may be characterized by the largest distance between their members, called their *diameter*. The technique minimizing the diameter of a cluster is variably called the complete linkage method (Sokal & Sneath, 1963), compact clustering (Lorr, cited in Cureton, Cureton, & Durell, 1970), or the diameter method (Johnson, 1967) and controls for what single-linkage clustering omits, namely, that extreme differences within a cluster stay within bounds. The largest distance within a cluster G , the diameter, is

$$d_{GG} = \max_{i,j \in G} (d_{ij})$$

The diameter of a cluster is a heterogeneity measure but it takes only one distance as representative of the cluster as a whole. In the formation of clusters that minimize this measure, a successive merging procedure will join at each step those two clusters whose most distant objects have the smallest distance across all pairs of clusters. Thus, by analogy to MacNaughton-Smith (1965) and Johnson (1967), if E and F are merged to form a new cluster G ,

$$d_{GG} = \min_{E, F} (d_{EF}) = \min_{E, F} \left(\max_{i \in E, j \in F} (d_{ij}) \right)$$

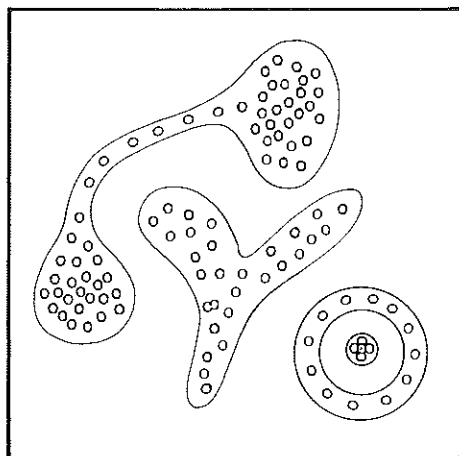


FIG. 12. Clusters for single-linkage method.

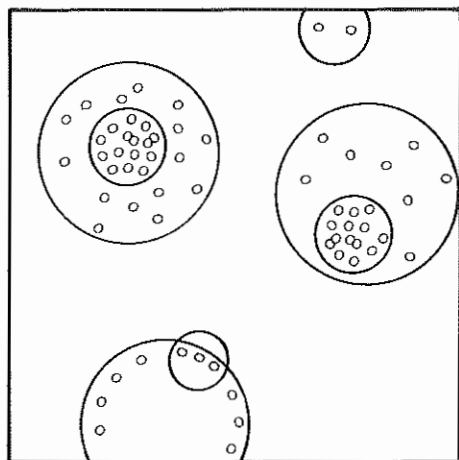


FIG. 13. Clusters for diametric method.

Compared with the connectedness criterion, the diameter criterion yields relatively compact clusters. But such clusters have several peculiar properties. First, there is the tendency of clusters to become circular and equal in diameter. Second, the number of objects within a cluster, the density of its population, has no bearing on the way clusters are formed. Third, and a corollary of the second, a center need not exist for such clusters. Figure 13 illustrates some of the typical clusters diametric clustering will yield. These clusters are shown at two stages of formation.

Average linkage clustering (Sokal & Sneath, 1963) extends the notion of connectedness to an aggregate measure of all distances between two clusters. So Bock (1974) defines the average distance between two clusters by

$$d_{E|F} = \frac{1}{n_E n_F} \sum_{i \in E} \sum_{j \in F} d_{ij}$$

The criterion to merge E and F into G , generalized to any power, then is

$$\min_{E, F} (d_{E|F}) = \min_{E, F} \left(\frac{1}{n_E n_F} \sum_{i \in E} \sum_{j \in F} d_{ij}^r \right)^{1/r}$$

Another average linkage criterion is Pearson's (1926) coefficient of racial likeness:

$$\min_{E, F} (d_{E|F}) = \min_{E, F} \left[\frac{1}{m} \sum_{u=1}^m \frac{(\bar{x}_{Eu} - \bar{x}_{Fu})^2}{(\sigma_{Eu}/n_E) + (\sigma_{Fu}/n_F)} \right]^{1/2}$$

in which \bar{x}_{Eu} is the arithmetic mean of values in E of variable u , and σ_{Eu} is the variance within E regarding the u th variable. The coefficient has been used by Rao (1948, 1952), Sokal and Sneath (1963), and several others. The coefficient makes the difference between two clusters a function of the

variance within both and is thus not a "pure" average linkage measure but belongs to the same family.

Both difference measures work against the chainlike appearance of clusters typical of the connectedness criterion but do not eliminate it completely. Their clusters are less compact than those using corresponding heterogeneity measures. Difference measures simply do not optimize homogeneity within a cluster. Once two clusters are merged, the distances that contributed to that decision are no longer referred to in subsequent clustering steps.

Clusters with minimum *variance* between their objects are achieved by taking all distances within a prospective cluster into account. Accordingly, a multiple linkage measure of heterogeneity *within* a cluster G is

$$d_{GG} = \left(\frac{1}{n_G(n_G - 1)} \sum_{i,j \in G} d_{ij}^r \right)^{1/r}$$

and to minimize this heterogeneity, clusters E and F are merged into G when

$$d_{GG} = \min_{E,F} (d_{EF}) = \min_{E,F} \left(\frac{1}{(n_E + n_F)(n_E + n_F - 1)} \sum_{i,j \in E,F} d_{ij}^r \right)^{1/r}$$

When the exponent $r = 1$, d_{GG} is the mean distance within the objects of a cluster and its use as a decision criterion assures that this mean distance is kept at a minimum. When $r = 2$, d_{GG} is the standard deviation within a cluster. And, since the variance is the square of the standard deviation, clustering with $r = 2$ also might be said to minimize the variance within clusters. Sokal and Sneath (1963) termed the latter index, the taxonomic distance.

Variance-type heterogeneity measures compensate for a large distance within a cluster by several smaller distances within that cluster. Clusters with equal heterogeneity in this variance sense may thus be different in diameter. However, as r increases in value, larger distances are weighted increasingly heavily on the measure so that r in fact controls the conservatism of the clustering procedure. The higher the exponent r the more compact the clusters that emerge. Some typical clusters that variance methods will identify are depicted in Fig. 14.

The relationship between variance-type difference and heterogeneity measures is easily illustrated by the following equality in which $r = 1$ for both the average linkage between clusters and the mean distance within the union of two clusters:

$$(n_E + n_F)(n_E + n_F - 1)d_{EF} = n_E(n_E - 1)d_{EE} + n_E n_F d_{E|F} + n_F(n_F - 1)d_{FF}$$

This equality reveals the difference measure $d_{E|F}$ to be *only one part* of the heterogeneity measure d_{EF} . If used as a clustering criterion, $d_{E|F}$ ignores the heterogeneities d_{EE} and d_{FF} of the clusters being merged and thus minimizes some property other than a characteristic of all the objects in the merging cluster. This accounts for what was suggested earlier, that difference measures

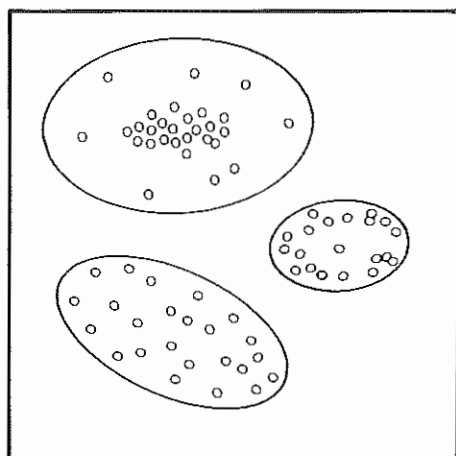


FIG. 14. Clusters for variance method.

tend to let clusters “grow out of control” and are for this reason inferior to measures of heterogeneity.

Clusters may also be *polythetic* or *monothetic* (Sokal & Sneath, 1963). In a monothetic cluster *all members share* some properties. Objects are admitted to a cluster because a large number of their characteristics match. In a polythetic cluster, members need not hold any value in common. They need only be similar in some respect and this similarity may be expressed by a high correlation, by a larger proximity between values, or by the sharing of values between pairs rather than among all members of a cluster. Distance recursive techniques yield polythetic clusters only and all clustering criteria discussed so far are also polythetic in result.

An example of a monothetic clustering technique is Krippendorff's (1975) strong associative clustering of binary attribute data. Key to the technique is the successive enumeration of the attributes that are shared among all members of a cluster. Since this number cannot be obtained from agreements between pairs, the enumeration must be data recursive. And seeking to correct observed agreement on attributes by what is due to chance, one of the more convincing coefficients turns out to be a generalization of Benini's (1901) measure of association. With a_G as the proportion of attributes shared among objects in G and e_i as the proportion of attributes associated with the object i of G , the coefficient, converted into a heterogeneity measure by $d_{ij} = 1 - s_{ij}$ and stated as a decision criterion for merging E and F into G , is

$$d_{GG} = \frac{\min_{i \in G}(e_i) - a_G}{\min_{i \in G}(e_i) - \prod_{i \in G} e_i} = \min_{E, F}(d_{EF}) = \min_{E, F} \left(\frac{\min_{i \in E, F}(e_i) - a_{EF}}{\min_{i \in E, F}(e_i) - \prod_{i \in E, F} e_i} \right)$$

It is the proportion of the observed disagreement on attributes shared within G to the disagreement of chance matching.

A graphical representation of typical clusters resulting from this monothetic, heterogeneity measure is difficult precisely because multiordinal clusters defy spatial representations. Further elaboration of the measure and an example are found in Section IX. This heterogeneity measure exemplifies a measure that does not satisfy the triangle inequality and would thus not lend itself to distance recursive clustering.

An example of a polythetic multiordinal and variance-type technique is multivariate classification. It is a method by which clusters are formed neither by grouping objects in terms of their variables nor by grouping variables in terms of their objects (all of which might produce one univariate classification or clustering scheme) but by clustering variables in terms of each other, interactionally so to speak, using the distribution of objects in multivariate space as a reference for the interaction. Krippendorff (1969, 1974) developed the technique from information theory. The simplest decision criterion is given in the section on distances. And the recursive form of the heterogeneity measure is presented in Section IX where details are elaborated. This measure assesses the amount of multivariate structure lost within the m -dimensional space when some of the terms within variables are no longer differentiated. It can be interpreted as expressing the amount of multivariate information that can no longer be transmitted due to the formation of clusters in each dimension or as the amount of relational entropy lost within a cross-classification of m separate clustering schemes, one for each variable. What this clustering technique achieves is a more efficient representation of the objects involved, one that reduces the m -dimensional space in volume without much loss in the essential relationship (see Figs. 20 and 21 for examples). Since the complexity of the resulting cluster again defies a simple graphical representation, Fig. 15 offers a two-dimensional diagram of the nature of the clusters that the technique might identify.

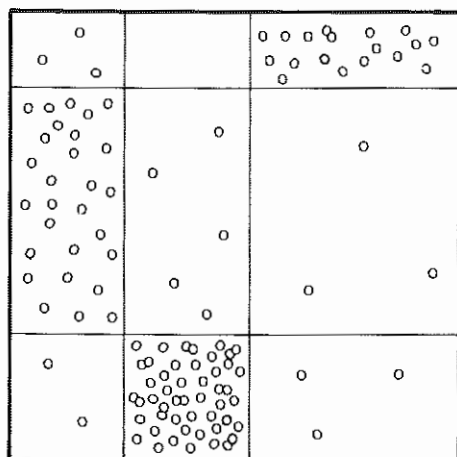


FIG. 15. Clusters for multivariate classification method.

In conclusion, it is evident that the choice of clustering criteria is the most important determinant of the kind of clusters that do emerge and the kind of properties these clusters are thereby able to represent. The classification of these properties merely serves to clarify principal differences between clustering criteria and the emerging properties of clusters. Examples are more numerous for difference than for heterogeneity measures, for single than for multiple linkage procedures, for polythetic than for monothetic clusters. The reliance of biordinal rather than multiordinal conceptions of properties is striking, and the fact that I did not exemplify the results of partitioning approaches to clustering is indicative of suspicious white spaces on the map of all combinatorially possible clustering techniques.

IX. CLUSTERING ALGORITHMS

An algorithm is a stepwise procedure that is completely specified (leaves no alternative undecidable) and transforms some input into some output. Clustering algorithms accept data as input either in their canonical form of a data matrix X or in their derived form of a distance matrix D . The output of a clustering algorithm either identifies a cluster containing a given object, produces some partition, i.e., a set of clusters, satisfying some criterion of optimality, or it gives a decision tree that contains partitions all of which satisfy some criterion of optimality.

Because of the insurmountable efforts required to compute clusters, partitions, and decision trees simultaneously, demonstrated previously, algorithms must be defined recursively. A recursive algorithm is one that is applied to some initial set of data, yielding an output to which it is applied again and again until some terminating criterion is met. Recursive clustering algorithms work themselves stepwise either down a partition lattice (through successive partitioning) or up the partition lattice (through successive merging). After the first step, a recursive algorithm avoids references to the initial data and transforms only its transforms.

This section presents four different algorithms for clustering objects or variables. The algorithms are chosen for their distinctive features: successive partitioning versus successive merging, biordinal clustering versus multiordinal clustering, data recursive versus distance recursive, monothetic versus polythetic, etc. The algorithms presented here do not exhaust all alternatives however. A researcher has many more options than are given here.

A. The Johnson Algorithm

The first example of a clustering algorithm is taken from Johnson (1967), who formalizes the two single-linkage clustering techniques previously discussed. Since both are distance-recursive merging techniques and proceed identically except for their clustering criteria, I shall describe only one here, the algorithm for diametric clustering. The Johnson algorithm is extremely

simple, speedy in execution, and therefore inexpensive to use, which might explain its widespread application.

Given: The $n \times n$ matrix of distances $(d_{ij}^s) = D^s$ at the initial $s = 0$.

Step 1 Search for the smallest distance, $\min_{E \neq F} (d_{EF}^s)$ in the $(n - s) \times (n - s)$ matrix D^s .

Step 2 Print or store on a history record: $s + 1$, E , F for which d_{EF}^s is minimum and merge all pairs of objects in E and F into G .

Step 3 Compute a new $(n - s - 1) \times (n - s - 1)$ matrix of distances in D^{s+1} from D^s by replacing all entries with references to E and F by

$$d_{GI}^{s+1} = \max(d_{EI}^s, d_{IF}^s) \quad \text{for all } I \neq E, F \text{ in } D^s$$

Step 4 Set $s \leftarrow s + 1$, return the new D^s to Step 1 unless either the smallest distance d_{EF}^{s-1} , now d_{GG}^s , exceeds a specified limit, the number of remaining clusters $n - s$ falls below a specified number, or any other terminating criterion is satisfied.

Step 5 Print results and terminate.

For an illustrative example, Johnson uses data obtained from a psycho-acoustic study of 16 principle consonants which are listed here across Fig. 16. The numbers down the left-hand side are similarity values that were obtained in the study and are here associated with each merging as indicated. Apparently the resulting clusters correspond to the distinctive features presumed by Miller and Nicely (1955) who provided these data. At the level of five clusters the 16 phonemes divide into a hierarchy, as depicted in Fig. 17.

	<i>p</i>	<i>t</i>	<i>k</i>	<i>f</i>	<i>θ</i>	<i>s</i>	<i>ʃ</i>	<i>b</i>	<i>d</i>	<i>g</i>	<i>v</i>	<i>ð</i>	<i>z</i>	<i>ʒ</i>	<i>m</i>	<i>n</i>
—
2.635	XXX
2.234	XXX	XXX
2.230	XXX	.	.	.	XXX	.	.	.	XXX
2.123	XXX	.	.	.	XXX	.	.	.	XXX	.	.	XXX
1.855	XXXXXX	.	.	.	XXX	.	.	.	XXX	.	.	XXX
1.683	XXXXXX	.	XXXXXX	XXX	.	.	XXX
1.604	XXXXXX	.	XXXXXX	XXX	.	.	XXX	XXX	.	.	.
1.525	XXXXXX	.	XXXXXX	XXX	.	XXXXXX	XXX
1.186	XXXXXX	.	XXXXXX	XXX	XXXXXXXX	XXX
1.119	XXXXXX	.	XXXXXX	XXXXXX	XXXXXX	XXXXXXXX	XXX
0.939	XXXXXX	XXXXXXXX	XXXXXX	XXXXXXXX	XXX	.	.	.	XXXXXXXX	XXX
0.422	XXXXXXXXXXXXXXXX	XXXXXX	XXXXXXXX	XXX	XXXXXXXX	XXX
0.302	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXX	XXXXXXXX	XXX
0.019	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXXXXXXX	XXX
0.000	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXX	XXXXXXXX	XXX

FIG. 16. Computer printable dendrogram for phoneme clusters (Johnson, 1967).

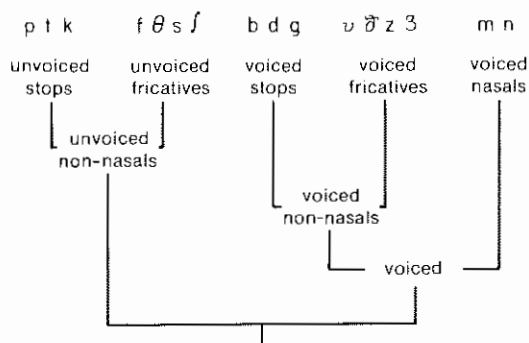


FIG. 17. Typology for phonemes derived from clustering (Johnson, 1967).

B. The CONCOR Algorithm

The second example of a clustering algorithm is taken from Breiger, Boorman, and Arabie (1975). It is the only effective partitioning algorithm I know that is applicable to a wide variety of data. CONCOR is the acronym of "convergent correlation" which designates the iterative application of correlations to yield stable indications of dependencies which are in turn used to partition a set of objects or variables into two parts. The use of correlations identifies the method as a biordinal partition technique.

Given: The $n \times m$ matrix $(x_{iu}) = X$ with interval metric in each variable u .

Step 1 Compute the $n_G \times n_G$ matrix of product-moment correlations $(r_{ij}^s)_G = {}_tR_G^s$ for the initial $s = 0$ and $r = 0$ where the initial set G is the set of all $n_G = n$ objects.

Step 2 Compute the $n_G \times n_G$ matrix of product-moment correlations ${}_{t+1}(r_{ij}^s)_G = {}_{t+1}R_G^s$ from all pairs i and j in ${}_tR_G^s$ and iterate this step to create from ${}_0R_G^0 = {}_0R_G^s$, ${}_1R_G^s$, ${}_2R_G^s$, ${}_3R_G^s$, ... until all entries r_{ij}^s of ${}_wR_G^s$ approximate within a specified limit the value $+1$ or -1 .

Step 3 Permute ${}_wR_G^s$ into the bipartite form:

	<i>E</i>	<i>F</i>
<i>E</i>	+1	-1
<i>F</i>	-1	+1

Step 4 Print or store on a history record: $s + 1$, the partition of G into E and F , and decompose the original correlation matrix ${}_0R_G^0$ into and store separately the two submatrices, the $n_E \times n_E$ matrix ${}_0R_E^0$ and the $n_F \times n_F$ matrix ${}_0R_F^0$.

Step 5 Set $s \leftarrow s + 1$, search for the largest remaining matrix ${}_0R_I^0$, and return that matrix as ${}_0R_G^s$ to Step 2 unless either s exceeds a specified limit or the number n_G of the largest cluster falls below a specified number.

Step 6 Print results and terminate.

The authors developed this algorithm for clustering a variety of sociometric data and profiles of either objects or variables between which correlations can be computed. This requires that all variables possess interval metric. There is no reason, however, to restrict the use of this algorithm to canonical forms of data and to variables with interval metrics. Since the principal feature of the algorithm is that correlations are computed iteratively, any matrix of distances or similarities with appropriate interval metric properties could be entered at Step 1 in place of ${}_0R_G^0$.

The authors have applied this algorithm to many sets of data and report that with $n = 70$ and the cutoff point for $r_{ij} = .999$, no more than 11 iterations are needed to approximate stability conditions. This keeps the required computational effort within practical limits. Apparently, while there are a few theoretical examples of a knife-edge character in which the iteration of correlations does not converge to the bipartite form, actual data that would lead to indecisions of this sort have not been encountered.

The interpretation of the CONCOR clusters is difficult however. As the authors recognize, the procedure does not use measures of homogeneity or heterogeneity as decision criteria, and while it is easy to understand when correlations are positive within clusters and negative across, such an understanding is indeed difficult when one is concerned with correlations of correlations . . . that might be 11 times removed from the data. Nevertheless Breiger, Boorman, and Arabie have compared CONCOR results with a variety of results obtained from other clustering techniques and found them convincing. Its clusters seem to be very similar to those obtained by clustering techniques based on a connectedness criterion.

C. The Strong Association Algorithm

In the next example I shall attempt to illustrate how a very simple multiordinal clustering procedure works and also how the required decision criteria may be formulated recursively to keep computational efforts small. It will be recalled that the computational effort of multiordinal clustering is generally magnified by the fact that such techniques require constant interaction between the procedure and data in their canonical form. Since clustering, to be practical, must proceed recursively, measures that keep account of the increasing heterogeneity in the emerging clusters should be defined recursively as well, else the procedure would have to return to the original data at each step and thereby annul the computational advantage of recursion.

The algorithm for strong association of 2^m data (Krippendorff, 1975) was developed in this way. It is applicable to binary attribute data where the

attributes to be shared within clusters are assigned the value $x_{iu} = 1$ and the absence of this attribute the value $x_{iu} = 0$. In terms of the 2×2 contingency matrix defined in Section V, Krippendorff's generalization of Benini's association coefficient, which is converted here to a suitable heterogeneity measure by $d_{ij} = 1 - s_{ij}$, is

$$d_{EF} = \frac{\min_{i \in E, F}(e_i) - a_{EF}}{\min_{i \in E, F}(e_i) - \prod_{i \in E, F} e_i}$$

where, in terms of the canonical form of data, $e_i = m^{-1} \sum_{u=1}^m x_{iu}$ is the proportion of attributes present in object i , $a_{EF} = m^{-1} \sum_{u=1}^m \prod_{i \in E, F} x_{iu}$ is the proportion of attributes shared within E and F , $\min_{i \in E, F}(e_i)$ is the largest possible proportion of attributes that could be shared within E and F , and $\prod_{i \in E, F} e_i$ is the expected proportion of attributes that would be shared if their co-occurrence were due to chance.

It turns out, all these components of the measure can be defined in terms of one matrix and two quantities for each cluster at the initial $s = 0$ and at any $s + 1$ from s . At $s = 0$ and from the initial $n \times m$ matrix X^0 with entries x_{iu} , the maximum proportion of attributes in cluster $\{i\}$, containing just one member, is

$$\mu_{(i)}^0 = \frac{1}{m} \sum_{u=1}^m x_{iu} = e_i$$

and so is the initial probability of attributes in that one-object cluster:

$$\rho_{(i)}^0 = \frac{1}{m} \sum_{u=1}^m x_{iu} = e_i$$

At each prospective merger of two clusters E and F into G , these quantities change as follows:

$$\mu_G^{s+1} = \min(\mu_E^s, \mu_F^s), \quad \rho_G^{s+1} = \rho_E^s \rho_F^s$$

And the $(n - s) \times m$ matrix X^s becomes the $(n - s - 1) \times m$ matrix X^{s+1} by

$$x_{Gu}^{s+1} = x_{Eu}^s x_{Fu}^s$$

So that the measure of heterogeneity for merging E and F into G becomes a function of values solely available at the preceding iteration:

$$d_{GG}^{s+1} = d_{EF}^s = \frac{\min(\mu_E^s, \mu_F^s) - m^{-1} \sum_{u=1}^m x_{Eu}^s x_{Fu}^s}{\min(\mu_E^s, \mu_F^s) - \rho_E^s \rho_F^s}$$

This recursive formulation provides the key to the following surprisingly efficient algorithm:

Given: The $n \times m$ matrix $(x_{iu}^s) = X^s$ at $s = 0$ with the presence of an attribute denoted by $x_{iu} = 1$ and its absence by $x_{iu} = 0$.

Step 1 For $i = 1, 2, \dots, n$ set $\mu_{(i)}^s = \rho_{(i)}^s = m^{-1} \sum_{u=1}^m x_{iu}^s$.

Step 2 Compute the $(n-s) \times (n-s)$ distance matrix $(d_{EF}^s) = D^s$ from the $(n-s) \times m$ matrix X^s by

$$d_{EF}^s = \frac{\min(\mu_E^s, \mu_F^s) - m^{-1} \prod_{u=1}^m x_{Eu}^s x_{Fu}^s}{\min(\mu_E^s, \mu_F^s) - \rho_E^s \rho_F^s}$$

Step 3 Search for $\min(d_{EF}^s)$ in D^s and for this smallest distance, print or store on a history record: $s+1$, $\min(d_{EF}^s)$, E , F , the newly assigned label G , and D^s if required.

Step 4 Compute those recursive accounts that are affected by the merger of E and F into G :

$$\mu_G^{s+1} = \min(\mu_E^s, \mu_F^s), \quad \rho_G^{s+1} = \rho_E^s \rho_F^s, \quad x_{Gu}^{s+1} = x_{Eu}^s x_{Fu}^s, \quad s \leftarrow s+1$$

Step 5 Return the reduced $(n-s) \times m$ matrix X^s to Step 2 unless either $\max(d_{GG}^s)$ exceeds a specified value or any other terminating criterion is satisfied.

Step 6 Print result and terminate.

The example given in Table VII starts with an initial data matrix X^0 describing 10 objects in terms of 16 variables. When the distribution of attributes are examined in this matrix, one may discover that the attributes of object j are fully contained in the attributes of the i th object, yielding, as it should, a distance of $d_{ij}^0 = 0$. Since monothetic clusters are represented by the attributes its objects share, the cluster $\{i, j\}$ then takes on the attributes i and j have in common, here those of j , which may be seen in the subsequent transform of the data matrix. Also objects e and f show the strongest possible association which $d_{ef}^0 = 0$ indicates. At the third iteration it is the objects g and h that are found least different. In terms of the 2×2 contingency table, the distance would be computed as follows:

		g	
		0	1
h	0	$\frac{6}{16}$	$\frac{2}{16}$
	1	$\frac{1}{16}$	$\frac{7}{16}$
		$\frac{9}{16}$	1

$$d_{gh} = \frac{\min(\frac{8}{16}, \frac{9}{16}) - \frac{7}{16}}{\min(\frac{8}{16}, \frac{9}{16}) - \frac{8}{16} \frac{9}{16}} = .29$$

d_{gh} of D^3 then appears in D^4 's diagonal as associated with the cluster $\{g, h\}$. So the process continues as indicated in Fig. 18.

TABLE VII

HISTORY OF TRANSFORMS OF DATA IN CANONICAL FORM, EXAMPLE

	X^s										D^s									
<i>a</i>	0	1	1	1	0	1	1	1	0	1	1	1	0	1	0		.00			
<i>b</i>	1	0	1	1	1	0	1	0	1	0	1	0	1	0	0	1	.96	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	0	0	1	1	0	1		1.16	.64	.00	
<i>d</i>	0	0	1	0	1	1	1	0	1	1	0	0	1	1	0		.71	.89	1.07	.00
<i>e</i>	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0		.83	.96	.87	.00
<i>f</i>	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	.36	1.19	1.42	.76
<i>g</i>	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1.20	1.00	1.60	1.14
<i>h</i>	1	1	0	0	1	0	1	0	0	0	1	1	1	0	1		1.42	1.19	1.42	1.27
<i>i</i>	0	1	0	1	1	1	0	1	0	0	1	0	1	1	1		1.28	1.60	1.28	1.19
<i>j</i>	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	.91	1.91	1.83	.98
<i>a</i>	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0	.00			
<i>b</i>	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1		.96	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1.16	1.00	.00	
<i>de</i>	0	0	1	0	1	1	1	1	0	1	1	0	0	1	1	0	.42	.68	.63	.00
<i>f</i>	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	.36	1.19	1.42	.54
<i>g</i>	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1.20	1.00	1.60	.82
<i>h</i>	1	1	0	0	1	0	1	0	0	0	1	1	1	0	1		1.42	1.19	1.42	.91
<i>ij</i>	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	.50	1.17	1.00	.56
<i>a</i>	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0	.00			
<i>b</i>	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1		.96	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1.16	.64	.00	
<i>de</i>	0	0	1	0	1	1	1	1	0	1	1	0	0	1	1	0	.42	.68	.63	.00
<i>f</i>	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	.36	1.19	1.42	.54
<i>gh</i>	1	1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	.82	.77	1.02	.80
<i>ij</i>	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	.50	1.17	1.00	.56
<i>af</i>	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	0	.36			
<i>b</i>	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1		.97	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1.05	.64	.00	
<i>de</i>	0	0	1	0	1	1	1	1	0	1	1	0	0	1	1	0	.45	.68	.63	.00
<i>gh</i>	1	1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	.64	.77	1.02	.80
<i>ij</i>	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	.56	1.17	1.00	.56
<i>ade</i>	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	.45			
<i>b</i>	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1		.67	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0	1	.68	.64	.00	
<i>gh</i>	1	1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	.92	.77	1.02	.29
<i>ij</i>	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	.79	1.17	1.00	.52
<i>ade</i>	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	.45			
<i>b</i>	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1		.67	.00		
<i>c</i>	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0	1	.68	.64	.00	
<i>ghij</i>	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	.88	.96	.98	.52
<i>ade</i>	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	.45			
<i>bc</i>	1	0	1	1	0	1	0	1	0	1	0	0	1	0	0	1	.75	.64		
<i>ghij</i>	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	.88	1.08	.52	
<i>abcde</i>	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	.75			
<i>ghij</i>	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1.01	.52		

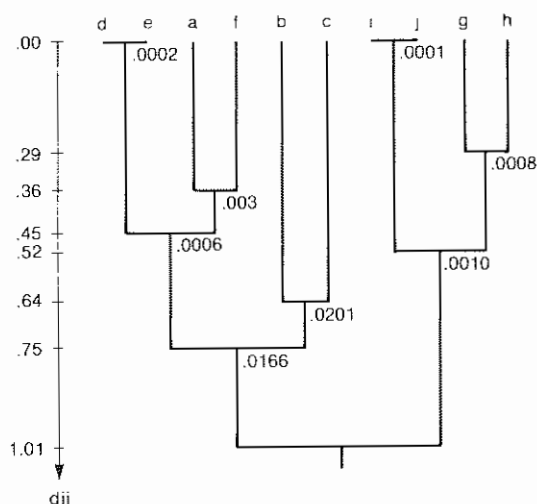


FIG. 18. Dendrogram for strong association clustering, an example.

The researcher now has to decide which partition is the most meaningful one, that is, at what level clusters are most convincingly interpretable. This is in part an intuitive decision but it can be strengthened by statistical considerations. In this example the null hypothesis (that the sharing of attributes within clusters is due to chance) can be rejected on significance levels indicated in the dendrogram at each point of merger. One can see that the significance level drops sharply after the sixth step and one might on these grounds be led to accept the partition of the ten objects into four clusters as optimal. However, the attributes then still overlap. A perfect differentiation between clusters is achieved only before the last step, at a point at which the two remaining clusters share no attributes anymore, with 7 out of the 16 variables providing the basis of the differentiation. All others dropped out.

D. The Multivariate Classification Algorithm

This example presents a clustering technique that does not cluster objects for their own sake, but rather uses them as a vehicle for simplifying their multivariate description. The description of these objects is a qualitative one; i.e., variables have the nominal metric throughout. Given a many-dimensional distribution of objects, the task of the clustering procedure is to reduce the representational space of this distribution not in dimensionality but in size without or with only a small amount of losses in structure within this space. Clusters then emerge not in one variable (e.g., the set of objects) and in terms of all other variables, rather, clusters emerge in all variables simultaneously, each in terms of all others. What is thereby taken account of is that the clustering within one variable may interact with the clustering in another variable and that higher-order dependencies within data are allowed to enter

such interactions. The most distinctive feature of this algorithm is that it optimizes the representation of multiordinal relations in data by the simultaneous clustering of values within many variables.

I am presenting here a version of the algorithm that is a considerable simplification of the one initially published (Krippendorff, 1974) and although multivariate classification provides several choices of heterogeneity measures, only the information theoretical measure of the amount of loss in structure will be used in the example. Since the procedure is a multiordinal one, and hence proceeds data recursively, the recursive formulation of loss functions is a key factor to the algorithm's practicality.

The algorithm yields several (as many as there are variables) hierarchical clustering schemes for the qualities in terms of which objects are described.

Given: The $n \times m$ matrix $(x_{iu}) = X$, all variables with nominal metric (unordered values).

Step 1 Reduce the $n \times m$ matrix X to an $(n - s) \times m$ matrix X^s containing $n - s$ unique objects i to each of which is assigned the frequency $n_{(i)}$. Compute frequencies $n_{x_u} = \sum_{i=1}^{n-s} n_{x_{iu}}$ for each value x occurring in variable u . (From here on X^s serves as a matrix of indices only.) Set $d_{(i)(i)}^s = 0$ for all $i = 1, 2, \dots, n - s$.

Step 2 With the function

$$\text{Loss}(a, b) = \begin{cases} 0 & \text{iff } a = b \text{ or } n_a = 0 \text{ or } n_b = 0, \\ (n_a + n_b) \log_2(n_a + n_b) - n_a \log_2 n_a - n_b \log_2 n_b & \text{otherwise} \end{cases}$$

with $n_{x_{Eu}}$ denoting the frequency of the value x_{Eu} within variable u in terms of which objects in cluster E are characterized, and with the m -valued description of each object divided into two parts, E_C and $K_{\bar{C}}$, so that $n_{\langle E_C K_{\bar{C}} \rangle}$ denotes the number of objects that share values x_{ic} within the set C of variables with objects in E but differ with respect to the remaining variables \bar{C} , now then compute the new $(n - s) \times (n - s)$ distance matrix D^s , replacing missing distances only, by

$$d_{EF}^s = d_{EE}^s + d_{FF}^s + d_{E|F}^s$$

where

$$d_{E|F}^s = \frac{1}{m} \left[\sum_C \sum_{K_{\bar{C}}} \text{Loss}(\langle E_C K_{\bar{C}} \rangle, \langle F_C K_{\bar{C}} \rangle) - \sum_{u=1}^m \text{Loss}(x_{Eu}, x_{Fu}) \right]$$

and where the sum over C refers to all subsets of variables in C whose values differ between E and F .

Step 3 Search for $\min_{E, F}(d_{EF}^s)$ in D^s and print or store on a history record $s + 1$, D^s if desired, $\min_{E, F}(d_{EF}^s)$, and for this minimum: E , F , the newly assigned label G , and for all values $x_{Eu} \neq x_{Fu}$: x_{Eu} , x_{Fu} , x_{Gu} , u .

Step 4 Merge E and F into G by modifying for all $x_{Eu} \neq x_{Fu}$ and u :

$$n_{x_{Gu}} = n_{x_{Eu}} + n_{x_{Fu}} \quad n_{x_{Eu}} = n_{x_{Fu}} = 0$$

for all clusters whose objects share some value x with values in $K_{\bar{C}}$ of E and F :

$$n_{\langle G_C K_{\bar{C}} \rangle} = n_{\langle E_C K_{\bar{C}} \rangle} + n_{\langle F_C K_{\bar{C}} \rangle}, \quad n_{\langle E_C K_{\bar{C}} \rangle} = n_{\langle F_C K_{\bar{C}} \rangle} = 0$$

and

$$d_{GG} = d_{EF}, \quad d_{EI} = d_{IF} = 0 \quad \text{for all } I \neq E, F$$

recompute s so that the number of unique objects or clusters is $n - s$.

Step 5 Return altered accounts to Step 2 unless either $\max(d_{GG})$ exceeds a specified limit, the number of clusters $n - s$ falls below a specified number, or any other terminating criterion is satisfied.

Step 6 Print results and terminate.

Note that d_{EF} measures the amount of structure lost by merging E and F . Within the three dimensions u , v , and w , if $x_{Eu} = x_{Fu}$ are the values E and F share, the distance between E and F expresses the difference it would make to the total amount of structure in the data when (with K denoting clusters other than E or F) all triples $\langle x_{Ku}, x_{Kv}, x_{Ew} \rangle$ and $\langle x_{Ku}, x_{Kv}, x_{Fw} \rangle$, $\langle x_{Ku}, x_{Ev}, x_{Kw} \rangle$ and $\langle x_{Ku}, x_{Fv}, x_{Kw} \rangle$, and $\langle x_{Ku}, x_{Ev}, x_{Ew} \rangle$ and $\langle x_{Ku}, x_{Fv}, x_{Fw} \rangle$ would no longer be differentiated. The sum over C then assures that all clusters are merged whose objects share some value with E or F in v , in w , and in both vw . What d_{EF} assesses is the effect of collapsing not only the point E and F but also all planes on which these points are located (see Fig. 19).

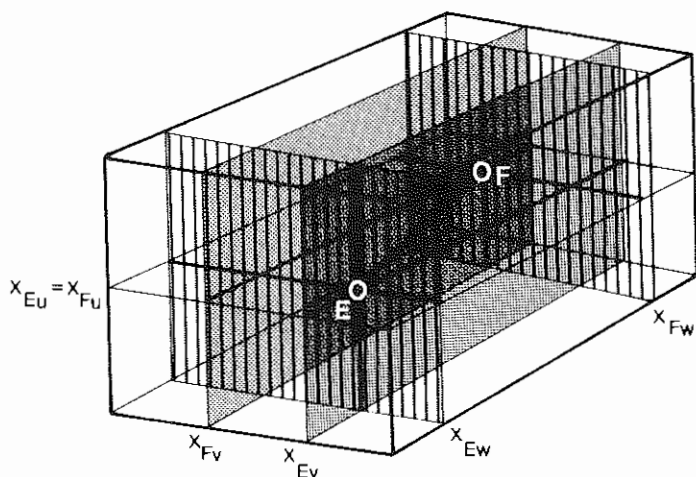


FIG. 19. A multivariate clustering step in three-dimensional space.

The implementation of the previously published version of the multivariate classification algorithm used an extremely wasteful form of storage (objects occupied cells in an m -dimensional array) and more complicated accounting devices that made the procedure reach computational limits before practical results could be obtained. The preceding algorithm is currently under investigation.

What multivariate classification accomplishes might best be illustrated graphically in Fig. 20. Suppose a three-valued characterization of a sample of objects finds all objects distributed as in the left space. There is a lot of redundancy in the values used for describing these objects and there is also some structure manifest in the distribution. Multivariate classification would now attempt to eliminate this redundancy by grouping variables in such a way that the remaining space contains as much of the initial structure as possible. In the illustration on the right of the original distribution no structure is lost. The algorithm boiled the initial representation down to its essentials.

In another, somewhat more artificial example, consider the schematic figure of a man as in Fig. 21. The clustering of values in the horizontal dimension first eliminates the duplication of columns, here due to symmetry, and yields the figure to the right of the original, showing no loss. The clustering of values in vertical dimension eliminates all duplication of rows and yields the figure below the original, showing no loss either. Clustering in both dimensions yields the resultant figure below and right of the original, also showing no loss in structure. (At this point it might be said that the example is misleading insofar as the algorithm does not recognize proximities between rows and columns which are important in Gestalt perception.) The original figure can be reconstructed from the figure below and right of the original by inverse application of the hierarchical clustering that emerged in each variable.

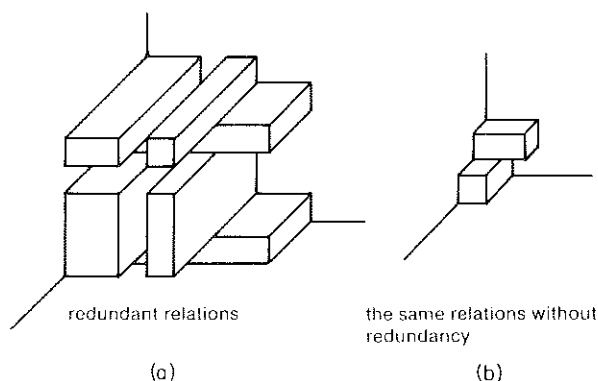


FIG. 20. Simplification of a distribution by multivariate classification. (a) Redundant relations; (b) the same relations without redundancy (Krippendorff, 1974).

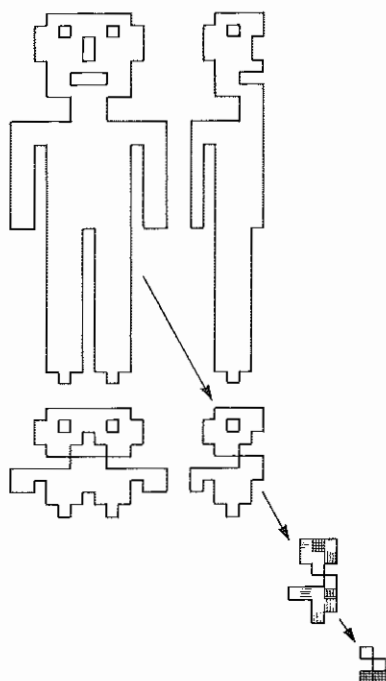


FIG. 21. Simplification of a figure by multivariate classification.

Further simplification of the figure results in losses which are indicated here by shading. When such losses occur they cannot necessarily be evaluated from either dimension in isolation. Losses in structure are losses in interaction effects and thus require that losses be inspected simultaneously for all variables involved. Continuing the classification one might reach the figure in the extreme lower right corner. The two final steps would wipe out the structure in this simple figure. Where to stop the clustering is a question of applying suitable termination criteria on the process.

It should be mentioned that a variety of clustering algorithms appear in Hartigan (1975) whose work was published after this was substantially completed.

X. VALIDATION AND VINDICATION

Clustering procedures compute clusters, often regardless of how strong the patterns permeate the data these clusters aim to represent. In any distribution of objects in space, even in an entirely random one, some objects are bound to be closer to each other or more similar than are others. Even when neighboring objects are approximately equidistant, the slightest inequality can provide the kick that starts a clustering sequence rolling. Testing for the

statistical significance of the clusters that do emerge in the process therefore is an important safeguard against attempts to interpret the results of a clustering process when an underlying pattern is spurious or does not exist in fact. The minimum requirement is to test one of two null hypotheses: that the co-occurring qualities shared within a cluster are due to chance or that the objects of different clusters are drawn from the same population.

But evidence for the statistical significance of clusters should be seen only as a prerequisite for entering into validity considerations. Only when null hypotheses can be rejected with confidence might one find empirically meaningful interpretations of clustering results. When clustering turns out to be due to chance or an artifact of the procedure, their potential validity may be soundly questioned.

Although I am quite aware of existing typologies for kinds of validation, let me focus here on only two types that Feigl (1952) termed validation and vindication. In the context of this application, *validation* is a mode of justification according to which the results of a particular analytical procedure are justified by showing the structure of that procedure to be derivable from general principles or theories that are accepted quite independently of the procedure to be validated, while *vindication* is a mode of justification that renders a particular analytical procedure acceptable on the grounds that its results lead to accurate predictions (to a degree better than chance) regardless of the details of the procedure. The rules of deduction and induction are essential to validation while the relation between means and particular ends provide the basis for vindication. In focusing on these two kinds of justification, I take for granted that the procedure is reliable, that successive clustering is order invariant, that distances and homogeneity measures satisfy required conditions, etc., all of which can be justified on logical grounds. I also take for granted that data are relevant in the sense considered earlier, for it is inconceivable that valid clusters can be obtained from irrelevant data.

Two not necessarily separate questions pertain to the *validation* of clustering procedures. First, exactly what features of objects are characterized when data enter the procedure in their derived form as distance or similarity measures, and are these features and the omission of others justifiable on theoretical or on empirical grounds? And second, exactly what does a clustering procedure optimize; which clustering criteria does it employ; and how do the measures that characterize the emerging clusters relate to a theory about how objects become associated, group themselves, or are clustered in reality?

Regarding the first question, it should be noted that there are great differences between how product-moment correlations, Euclidean distances, or information losses conceptualize and quantitatively assess dissimilarities between objects. For example, the product-moment correlation assesses the degree to which two objects are linearly related. A positive r_{ij} indicates that the values of two objects increase in the same direction, while a negative r_{ij}

indicates such an increase to be in the opposite direction. But $r_{ij} = +1$ does not imply $i = j$. The underlying concept of resemblance is a very peculiar one, and the researcher who wishes to cluster on the basis of a correlation matrix must establish that available knowledge about the nature of the objects would indeed lead to this conception. Product-moment correlations assume that objects i and j are related by $a_j x_{ju} + b_j = a_i x_{iu} + b_i$ for all variables u and that similarity is independent of the constants a and b . Obviously this does not conform to intuition, which suggests that two objects are maximally similar only when $i = j$, that is, $x_{ju} = x_{iu}$ for all of the u . Pearson's intraclass correlation coefficient satisfies this condition, the product-moment coefficient does not.

An examination of whether the formal properties of the underlying distance or similarity measures are defensible ought to be made before any clustering for a particular purpose is undertaken. Failure to provide such validating evidence makes it otherwise difficult to interpret findings in the light of a given theory.

One common way of bypassing the validation of distances is to input data in the form of subjective difference or similarity ratings obtained from a sample of subjects. If the researcher is indeed interested in clustering subjective difference or similarity judgments, two problems tend to arise: One is the variance that such judgments invariably entail and the other is that such judgments ought to satisfy the formal conditions of a distance.

Answers to the second question pertaining to the validation of clustering criteria pose even greater problems. To decide on the acceptability of a clustering criterion, the researcher must first decide on the properties his clusters are to represent. Given such a designation of purpose the researcher must then examine the principles underlying the formation of the groupings in reality that a clustering attempts to approximate or predict. In order to complete the validation, the researcher must finally demonstrate consistency between the decision criteria, difference or heterogeneity measures employed in the clustering procedure on the one side and knowledge about the natural processes on the other. This knowledge may take the form of an established theory, of hard empirical evidence, or in its weakest form, of grounded intuition. Wherever such knowledge comes from, validation may rely on it.

For example, if the resulting clusters are expected to predict how individuals form cliques or other social forms of organization, then knowledge about the way such social groupings emerge is indispensable in the validation of a procedure. The knowledge that cliques and social groups possess synergetic-organizational-Gestalt qualities and that their formation cannot be predicted from information on the interaction within pairs of individuals would render biordinal techniques invalid from the start (unless the effect were insignificantly small). To be valid, a multiordinal technique would then have to replicate the social process involved.

Another crucial option is whether clusters are formed on the basis of differences between clusters or heterogeneities within. The chainlike clusters

resulting from the connectedness method have already been contrasted with the compactness of the diameter method. The knowledge that all similarities or distances within a cluster will determine its boundaries with other clusters lends validity to a minimum heterogeneity criterion. In contrast, the knowledge that significant similarities and distances follow a hierarchical pattern, representing differences between clusters while neglecting those within, would lend validity to a minimum difference criterion. A hypothetical example of a situation in which a difference measure might be superior to a heterogeneity measure would be a certain form of communication within a social organization in which communication occurs primarily *on the same level* of the organizational hierarchy and *between minimally different* parts of the organization and is secondary or absent *across different levels* of such a hierarchy and *within* these parts. Such an implicit hierarchical conception of difference would be inappropriate when an organization is formed on the basis that members share certain properties, that communication within is larger or more important than communication across the parts of an organization, etc.

It is often more difficult to apply available evidence about natural groupings on a given clustering technique than to formalize such evidence into a computable clustering criterion. The development of the strong association technique by Krippendorff (1975) is a case in point. It started with a problem in content analysis where the development of emic or indigenous as opposed to etic or imposed categories is a common problem. The task was to develop a reliable coding instrument for advertising appeals in categories that are close to those used by television viewers. For this purpose Dziurzynski (1978) asked subjects to group about 300 appeals culled from commercials into categories that seemed most meaningful to them. In observing the subject's justifications one often finds some like this: "If i and j are together, then k must be in the same category, but if h and i are in the same category then k cannot join them." Those are typical *multiordinal* arguments. The task was to form clusters among aspects based on *agreements* among subjects regarding the grouping. The formalization of the notion of agreement, which ought to be maximum when groups are either identical or when one is included in the other, leads to an association coefficient which has been discussed in the preceding. The correspondence was taken as validating evidence.

To summarize, validation asks whether the way information is processed within a clustering procedure is consistent with the way such information would be processed in the real world, while vindication asks the a posteriori question of whether the results of a clustering procedure correspond with independently obtained evidence about clusters.

The most obvious form of *vindication* is to establish correspondence between the results of a clustering procedure and the results obtained by other methods (including by independent observation). Since clusters obtained by other methods must always be available for such comparisons, vindication primarily yields information about the efficiency or simplicity and only secondarily about the adequacy of the underlying structure.

So, when developing the partitioning algorithm CONCOR and probably because correlations of correlations of . . . is a concept that is far removed from penetrations by intuition, Breiger *et al.* (1975) compared their results with those obtained by a variety of other clustering techniques. The finding that CONCOR results approximate those obtained by the connectedness method makes the procedure vindicatively acceptable but only to the extent the results of this connectedness method are already known to be valid in a particular application.

In another example of vindication, we asked subjects to group sets of words according to perceived semantic similarities. Since the multivariate classification algorithm was developed by formalizing certain theories of contextual meanings, if the theories and their algorithmic implementation are correct, then computational results and subjective clusters are expected to be in high agreement. In this case we were fortunate to be able to vary certain computational parameters and found, to our surprise, that the weakest clustering criterion resulted in the best fits. This may serve as a warning against the assumption that the validity of clustering procedures increases with their complexity.

In vindication experiments, the variability of a clustering technique is a deceptive virtue, however, for it is always possible to find a computational approximation to an independently obtained set of clusters. This possibility is exemplified in work done by Lance and Williams (1967) who showed with Fig. 22 how changes in value of one variable of their clustering criterion causes extremely different dendrograms to emerge from the same data. The danger is that once one considers oneself free to play with the clustering criteria, one can "prove" anything, and since the computation then merely supports what is already known, the proof given is "empty."

Vindication allows all conceivable clustering options to be tested against empirical evidence, but its aim is to find that option which produces *consistently* high agreements within the empirical domain chosen. A single "convincing match" means very little. Carefully used, vindication provides a method for generalizing or for confining the success of a particular clustering procedure.

The researcher who does not have independently obtained clusters at his disposal might be led to believe that the computational results "make sense" or are "acceptable on intuitive grounds." But a better way of rendering such results plausible is to get into the very procedure that produced them and to show that the procedure is, at least ideally, a homomorphic representation of the processes known to explain the phenomena under consideration. All users of clustering techniques should be expected to make at least some effort at validation when publishing their results.¹

¹Referring to the comments on this section by Tukey (see Chapter 16, Section B), I disagree that validation is impossible or dangerous but I am perfectly happy with his words: "All users of clustering techniques should be expected to make at least some effort (a) to explain why they

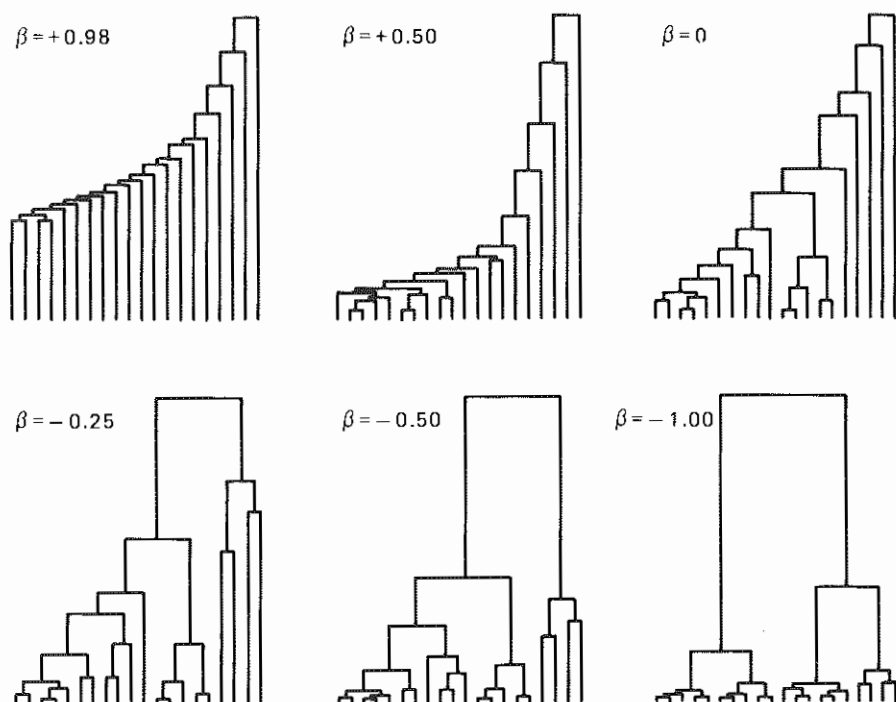


FIG. 22. Results of applying six different clustering criteria on the same objects (Lance & Williams, 1967, with permission of British Computer Society).

XI. SUMMARY—CONCLUSION

This chapter explores clustering as a multivariate technique in communication research and does so with several kinds of users in mind.

There is, first, the researcher who wants to make use of data stemming from clustering, either for a secondary analysis by different techniques, for supporting practical decisions, or simply, to understand published findings. If he seeks a level of understanding beyond the Presentation of Results he will want to acquaint himself with Validation and Vindication, needs to be able to

choose the methods used and (b) to make as clear as is reasonable the tentative character of clustering results in general and the degree to which this applies to those they describe."

This section is merely intended to put into focus the fundamental relationship between the description of objects, the process of clustering and its results (all of which are very much guided by the researcher's choices), and the nature of the objects and the processes by which objects form groups, cliques, classes, lumps, associations, or Gestalts (which are not so much influenced by if not independent of the way they are analyzed). No multivariate technique can avoid some degree of artificiality and its results are, hence, always tentative to some extent. The task of validating an analytic technique is to justify and to explain the use of a procedure not in terms of aesthetics, convenience, or habit but in reference to knowledge about reality, however hypothetical this might be.

judge the Relevance and Ordinality of Data from which results are obtained and read particularly the section on Properties of Emerging Clusters, at least where it pertains to the clustering procedure actually used. He may then be able to judge whether given clustering results may be interpretable in view of his particular problem.

There is, second, the researcher who seeks to apply one of the available clustering procedures to his data, with the aim of data simplification, in search for a typology or to group or lump together phenomena that share certain characteristics. Such a user may need to know the form of data amenable to clustering: Canonical Form of Data and Derived Form of Data. He may want to become familiar with the Presentation of Results. And, after familiarizing himself with the basic ideas of Validation and Vindication, he may want to read all that needs to be known to understand what available clustering procedures do: Properties of Emerging Clusters, Ordinality of Data, Clustering Algorithms, etc. He may then be able to make intelligent choices among available procedures or find that the tasks he set for himself cannot be accomplished by Clustering.

There is, third, the researcher who wants to design his own special purpose clustering technique. Whether he is a computer programmer himself or delegates the writing of such a procedure to someone else, he ought to consider the warnings in Goals and Computational Efforts seriously before conceptualizing a Clustering Algorithm, taking most of the sections of this chapter and references to additional literature into account.

The chapter will be useful, fourth, to the computer programmer who will have to converse with empirically oriented social scientists when helping him either to implement, modify, or to develop anew suitable clustering procedures. Much too often have I found that differences in technical discourse prevent the full utilization of available analytical or intellectual resources. Computer programmers may be keenly aware of Goals and Computational Efforts and the nature of Clustering Algorithms but often lack understanding of the philosophical issues raised in Validation and Vindication and the special demands made by available social theory on the Properties of Emerging Clusters.

The chapter is on clustering. But several important issues point beyond this (here welcome) restriction, for example, the issue of validating the logic of an analytical procedure as opposed to merely vindicating its result or the issue of the ordinality in data and the ordinality a procedure can take. It is amazing that most current *multivariate* techniques are *biordinal* in structure and thus fail to deliver what their label seems to suggest. So far we have always thought in categories: variance analysis, multidimensional scaling, clustering, etc., each had its own purpose and assumptions. Ultimately these categorical distinctions need to be overcome by tying the processes they follow more directly to those of the empirical world. These issues are of concern, finally, to the methodologist and epistemologist of the social sciences.

References

- Bailey, K. D. Cluster analysis. In D. R. Heise (Ed.), *Sociological Methodology 1975*. San Francisco, California: Jossey-Bass, 1975.
- Benini, R. *Principii di demografia*. No. 29. Manuali Barbera de Scienze Guiviche Sociale e Politiche, Firenze: Barbera, 1901.
- Bock, H. H. *Automatische Klassifikation*. Gottingen: Vandenhoeck and Ruprecht, 1974.
- Breiger, R. L., Boorman, S. A., and Arabie, P. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 1975, **12**, 328–383.
- Cureton, H. R., Cureton, L. W., and Durell, R. C. A method of cluster analysis. *Multivariate Behavioral Research*, 1970, **5**, 101–106.
- Driver, H. E., and Kroeber, A. L. Quantitative expression of cultural relationships. *University of California Publications in American Archeology and Ethnology*, 1932, **31**, 211–256.
- Dziurzynski, P. S. *Development of a content analytic instrument for advertising appeals used in prime-time television commercials*. MA Thesis. Philadelphia: The Annenberg School of Communication, University of Pennsylvania, 1978.
- Feigl, H. Validation and vindication: An analysis of the nature and limits of ethical arguments. In W. Sellars and J. Hospers (Eds.), *Readings in ethical theory*. New York: Appleton, 1952.
- Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrics*, 1966, **53**, 325–338.
- Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics*, 1971, **27**, 857–872.
- Gower, J. C., and Ross, G. J. S. Minimum spanning trees and single-linkage cluster analysis. *Applied Statistics*, 1969, **18**, 54–64.
- Hamming, R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, 1950, **26**(2), 147–160.
- Hartigan, J. A. *Clustering algorithms*. New York: Wiley, 1975.
- Heinecke, F. Naturgeschichte des Herings. I. Die Lokalformen und die Wanderungen des Herings in den Europäischen Meeren. *Abhandlungen des Deutschen Seefischerei-Vereins*, 1898, **2**, i-cxxvi, 1–223.
- Jardine, N. and Sibson R. *Mathematical taxonomy*. New York: Wiley, 1971.
- Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, **32**, 241–254.
- King, B. Step-wise clustering procedures. *Journal of the American Statistical Association*, 1967, **62**, 86–101.
- Krippendorff, K. *Computer programs for multivariate classification in content analysis, A proposal to the national science foundation*, Philadelphia, Pennsylvania: University of Pennsylvania, mimeo, 1969.
- Krippendorff, K. *Reliability*. Philadelphia, Pennsylvania. Univ. of Pennsylvania, mimeo, 1973.
- Krippendorff, K. An algorithm for simplifying the representation of complex systems. In J. Rose (Ed.), *Advances in cybernetics and systems*. New York: Gordon and Breach, 1974.
- Krippendorff, K. *A method for strong associative clustering of 2^m data*. Philadelphia, Pennsylvania: University of Pennsylvania, mimeo, 1975.
- Krippendorff, K. A spectral analysis of relations. Unpublished paper presented to the International Congress of Communication Sciences, Berlin, May 31, 1977. Philadelphia: The Annenberg School of Communications, University of Pennsylvania, mimeo, 1976.
- Krippendorff, Klaus. *On the algorithm for identifying structures in multivariate data, including structures with loops*. Philadelphia: The Annenberg School of Communications, University of Pennsylvania, mimeo, 1980.
- Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies I. Hierarchical systems. *Computer Journal*, 1967, **9**, 373–380.
- Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies II. Clustering systems. *Computer Journal*, 1967, **10**, 271–277.

- Lorr, M. A. A review and classification of typological procedures. Paper read at the meeting of the American Psychological Association, San Francisco, California, 1968.
- Lysenko, O., and Sneath, P. H. A. The use of models in bacterial classification. *Journal of General Microbiology*, 1959, **20**, 284-290.
- MacNaughton-Smith, P. Some statistical and other numerical techniques for classifying individuals. *Home Office Research Unit Report No. 6*, London: H. M. Stationary Office, 1965.
- Mahalanobis, P. C. On the generalized distance in statistics. *Proceedings of the National Institute of Science India*, 1936, **2**, 49-55.
- Miller, G. A., and Nicely, P. E. An analysis of perceptual confusion among some english consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 338-352.
- Osgood, C. E., Suci, J. C., and Tannenbaum, P. H. *The measurement of meaning*, Urbana, Illinois: Univ. of Illinois Press, 1957.
- Pearson, K. On the coefficient of radical likeness. *Biometrika*, 1926, **18**, 105-117.
- Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society*, 1948, **B10**, 159-193.
- Rao, C. R. *Advanced statistical methods in biometric research*, New York: Wiley, 1952.
- Sneath, P. H. A., and Sokal, R. R. *Numerical taxonomy*, San Francisco, California: Freeman, 1973.
- Sokal, R. R., and Sneath, P. H. A. *Principles of numerical taxonomy*. San Francisco, California: Freeman, 1963.
- Tryon, R. C. *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor, Michigan: Edwards Brothers, 1939.
- Tyron, R. C., and Bailey, D. E. *Clustering analysis*. New York: McGraw-Hill, 1970.
- Tukey, J. W. Personal communication, 1977.
- Zubin, J. A. A technique for measuring likemindedness. *Journal of Abnormal & Social Psychology*, 1938, **33**, 508-516.