CLUSTER ANALYSIS

MARK S. ALDENDERFER
Northwestern University

ROGER K. BLASHFIELD
University of Florida, Gainesville

1. INTRODUCTION

An Ancient Chinese Classification of Animals

Animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, and (n) those that resemble flies from a distance (Jorge Luis Borges, *Other Inquisitions: 1937-1952*).

Classification is a basic human conceptual activity. Children learn very early in their lives to classify the objects in their environment and to associate the resulting classes with nouns in their language. Classification is also a fundamental process of the practice of science since classificatory systems contain the concepts necessary for the development of theories within a science.

"Cluster analysis" is the generic name for a wide variety of procedures that can be used to create a classification. These procedures empirically form "clusters" or groups of highly similar entities. More specifically, a clustering method is a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups.

Clustering methods have been recognized throughout this century, but most of the literature on cluster analysis has been written during the past two decades. The major stimulus for the development of clustering methods was a book entitled *Principles of Numerical Taxonomy*, published in 1963 by two biologists, Robert Sokal and Peter Sneath. Sokal and Sneath argued that an efficient procedure for the generation of biological classifications would be to gather all possible data on a set of

organisms of interest, estimate the degree of similarity among these organisms, and use a clustering method to place relatively similar organisms into the same groups. Once groups of similar organisms were found, the membership of each group could be analyzed to determine if they represented different biological species. In effect, Sokal and Sneath assumed that "pattern represented process"; that is, the patterns of observed differences and similarities among organisms could be used as a basis for understanding the evolutionary process.

The literature on cluster analysis exploded after the publication of the Sokal and Sneath book. The number of published applications of cluster analysis in all scientific fields has doubled approximately once every three years from 1963 to 1975 (Blashfield and Aldenderfer, 1978b). This rate of growth is much faster than that of even the most rapidly growing disciplines, such as biochemistry. There are two reasons for the rapid growth of the literature on cluster analysis: (1) the development of high-speed computers and (2) the fundamental importance of classification as a scientific procedure. Before computers, clustering methods were cumbersome and computationally difficult when applied to the large data sets in which most classifiers were interested. To cluster a data set with 200 entities requires searching a similarity matrix with 19,900 unique values. To search manually through a matrix of this size is a staggering and time-consuming task that few researchers (or their unlucky assistants) are willing to undertake. With the widespread availability of the computer, the time-consuming process of handling large matrices became feasible.

The second reason for the growth of interest in clustering is that all sciences are built upon classifications that structure their domains of inquiry. A classification contains the major concepts used in a science. The classification of elements, for instance, is the basis for understanding inorganic chemistry and the atomic theory of matter; the classification of diseases provides the structural basis for the science of medicine. Since clustering methods are viewed as objective, easily replicable ways to construct classifications, they have enjoyed widespread popularity across seemingly diverse scientific disciplines.

The social sciences have long maintained an interest in cluster analysis. Among the earliest of these studies were those by anthropologists who defined homogeneous culture areas by using methods of matrix manipulation (Czekanowski, 1911; see also Driver, 1965; Johnson, 1972). In psychology, cluster analysis was viewed as a "poor man's factor analysis" by one of its major proponents (Tryon, 1939). Other disciplines, most notably political science, were also involved in the early development of clustering in the social sciences. Although many of

the theories and applications that served as the basis for clustering in the past have been repudiated by later generations of scholars, all social sciences now have strong modern traditions in the use of clustering methods.

Despite their popularity, clustering methods are still poorly understood in comparison to such multivariate statistical procedures as factor analysis, discriminant analysis, and multidimensional scaling. The social sciences literature on clustering reflects a bewildering and often contradictory array of terminology, methods, and preferred approaches. Published guidance for the novice is sparse, and this, combined with the diversity of terminology and methodology, has created a complex maze that is virtually impenetrable. The goal of this book is to guide the novice through the maze of cluster analysis. Because of the tremendous diversity of methods that have been proposed over the past twenty years, we do not exhaustively review all or even most of these methods. Instead, we emphasize those methods and procedures that are comparatively well known in the social sciences or that we believe have strong merit for use in applied research.

How Clustering Methods Are Used

As we have already noted, clustering methods are designed to create homogeneous groups of cases or entities called clusters. Most of the varied uses of cluster analysis can be subsumed under four principal goals:

(1) development of a typology or classification,
(2) investigation of useful conceptual schemes for grouping entities,
(3) hypothesis generation through data exploration, and
(4) hypothesis testing, or the attempt to determine if types defined through other procedures are in fact present in a data set.

Of these goals, the creation of classifications probably accounts for the most frequent use of clustering methods, but in most cases of applied data analysis, many of these goals are combined to form the basis of the study. To understand these goals better, consider the following illustrations of the use of cluster analysis.

Alcoholism is a major mental health problem in this country, but no classification of alcoholics has ever gained wide acceptance among mental health professionals. Goldstein and Linden (1969), two clinical psychologists, used cluster analysis to build a classification of alcoholics. They gathered data on 513 alcoholics who had been admitted to a rehabilitation program at a state hospital in Indianapolis, Indiana. The

data gathered on these patients were from a commonly used psychological test, the Minnesota Multiphasic Personality Inventory (MMPI). This test contains 566 true/false items that are standardly summarized in terms of 13 scales that have diagnostic significance (e.g., the Schizophrenia scale and the Hysteria scale).

Goldstein and Linden subdivided their data into two parts: a derivation subsample (239 patients) and a replication subsample (251 patients). Using the derivation subsample, they formed a 239 × 239 correlation matrix that represented the similarities among the MMPI profiles from these patients, and used a clustering method devised by Lorr (1966). Of the patients in the derivation subsample, 114 were assigned to four clusters, while the remaining 125 patients were not assigned to any cluster. When the same steps were performed on the replication subsample, four clusters were again found that contained 106 (of the 251) alcoholics. The mean profiles of the clusters were basically the same across the two subsamples. The names that Goldstein and Linden assigned to these four clusters of alcoholics were (1) emotionally unstable personality, (2) psychoneurotic with anxiety/depression, (3) psychopathic personality, and (4) alcoholic with drug abuse and paranoid features.

The Goldstein and Linden study was important in the literature on alcoholism because it provided the model for more than 15 subsequent studies using cluster analysis to classify alcoholics. Most of these studies have provided general support for the validity of the first two clusters (types I and II).

The second study examined was performed by two anthropologists, Burton and Romney (1975). Their goal was to determine on what basis speakers of English classified role terms. The data used in the study were the results of a simple sorting of the 58 most common role terms in the language. Typical of the terms included in the study were "artist," "boss," "friend," "man," "owner," "poet," and "spy." Those who participated in the study were given each of these terms on a separate piece of paper and asked to look at the terms and then to place them into whatever groups came to mind. No limits were placed on the number or size of the groups. The similarity between the groups of role terms was computed using the Z measure, commonly used with sorting-task data (Miller, 1969).

The authors first investigated the similarity data with nonmetric multidimensional scaling in an attempt to determine if any underlying structure could be used to describe the patterns of similarity of the role terms. A three-dimensional solution was deemed valid. The dimensions were interpreted as an evaluative dimension in which terms such as

"gambler," "gunman," and "spy" were contrasted with terms such as "friend" and "companion"; a power dimension in which formal roles such as "boss" or "foreman" were contrasted with kin terms and other intimate terms such as "friend"; and an occupational dimension that simply contrasted job roles with all other role terms. They then performed a hierarchical cluster analysis using two different methods on the similarity data. The authors chose the eight-cluster solution for each of the methods, and noted that the results from the two clustering methods, while different in many respects, nevertheless had four major clusters in common: (1) a cluster of seven kin terms, (2) a cluster of friends, (3) a cluster of social-category membership terms, and (4) a cluster of managerial roles. They concluded that the results obtained from the two multivariate methods were complementary, and suggested that those individuals who sorted the terms made their decisions both on the global criteria recovered by the multidimensional scaling (the dimensions of evaluation, power, and occupation) and the more finely tuned criteria suggested by the clustering of terms, such as the clear hierarchical structure of English kinship terms based upon degree of relationship between individuals regardless of sex. The results of the cluster analysis corroborated the ambiguity of sex roles in Western society that has been reported in other anthropological studies, and further clarified the basis upon which speakers of English classify kinship terms.

The final example, a sociological study by Filsinger, Faulkner, and Warland (1979), was designed to create a classification of religious individuals. The data were gathered using the Religiosity Scale (DeJong et al., 1976) which was administered in a questionnaire format to 547 undergraduate students at Pennsylvania State University. A total of 37 items were chosen from the Religiosity Scale, a measurement device based upon a previous factor analysis of these data (DeJong et al., 1976). Since the entire sample of 574 students was too large to analyze economically, a sample of 220 were selected for the study. A 220 × 220 similarity matrix between pairs of individuals was clustered. The authors chose the seven-cluster solution for interpretation, and named seven types of religious individuals:

Type I: outsiders
Type II: conservatives
Type III: rejectors
Type IV: moderately religious
Type V: marginally religious
Type VI: orthodox
Type VII: culturally religious

Filsinger et al. also attempted to validate their classification. First they performed a discriminant analysis on the clusters, and the results were said to be highly significant.[1] Second, they compared the subjects in the various clusters by using seven demographic variables. On four of the seven variables, (size of home community, political identification, percentage of students not associated with the church, and religious affiliation), the clusters were significantly different. The authors concluded that overall results provided support for their empirical typology of religious individuals.

Each of the goals of cluster analysis can be found in the examples. The building of classifications was the most important goal of the Goldstein and Linden, and Filsinger et al. studies, but the exploration of the classification schemes (the MMPI and the Religiosity Scale) figured prominently as well. The study by Burton and Romney was devoted primarily to data exploration and hypothesis testing, and the building of a formal classification was of secondary importance. In this case, while the hypothesis testing was not conducted formally, the authors observed that the results corroborated findings about the use of language that had been discovered through more traditional anthropological methods.

A consideration of the three examples also shows that despite differences in goals, data types, and methods used, five basic steps characterize all cluster analysis studies.

(1) selection of a sample to be clustered
(2) definition of a set of variables on which to measure the entities in the sample
(3) computation of the similarities among the entities
(4) use of a cluster analysis method to create groups of similar entities
(5) validation of the resulting cluster solution

Each of these steps is essential to the use of cluster analysis in applied data analysis, and each is discussed at length below.

Data Sets to Be Used as Examples

We will use one data set to illustrate how clustering methods are used, and one is presented in the Appendix. It is included so that interested readers can experiment with the procedures we illustrate; our results can be used as benchmarks for comparison.

The first example data set is a hypothetical collection of human burials and their accompanying artifacts from an archaeological site. Burial data are important to archaeologists because they may contain information about the social statuses or roles played by the individuals found in the graves. By carefully analyzing the contents of graves, archaeologists may be able to infer the status differences among individuals that in turn may lead to inferences about the nature of social ranking and level of development of the society that created them.

The data set varies across three dimensions: age, sex, and status. At our hypothetical archaeological site 25 individuals have been "interred," and they are divided into three age groups: children, adolescents, and adults. Two statuses are also present: elite and nonelite. Each grave may contain up to eight different artifactual types: local ceramics, arrow points, shell bead bracelets, chipped stone, bone pins, bone awls, imported ceramics, and metal. Each of these artifact types has a status and sex distribution; age distinctions in artifact types have not been included in the data so that the structure of the data set can remain relatively simple. The data are coded in binary form, with the simple presence or absence of the artifact recorded.

The second data set, also based upon artificial data, has been structured to represent the type of classification problem that is often addressed in psychopathology. The basic data set contains 90 hypothetical patients who represent three types of mental disorders: psychoses (P), neuroses (N), and character disorders (CD). Thirty patients were created to represent each of these general groups. Details of the data generation process can be found in Blashfield and Morey (1980). The variables used to assess the patients were the 13 standard scales from the MMPI, the psychological test described earlier in the Goldstein and Linden (1969) study of alcoholics. The names for these scales, with the abbreviations for each, are

- Validity Scales
  - L—Lie scale
  - F—"Fake bad" scale
  - K—Correction scale
- Clinical Scales
  - Hs—Hypochondriasis
  - D—Depression
  - Hy—Hysteria
  - Pd—Psychopathic Deviate
  - Mf—Male/Female scale
  - Pa—Paranoia
  - Pt—Psychasthenia
  - Sc—Schizophrenia
  - Ma—Hypomania
  - Si—Social Introversion

Briefly, the MMPI consists of 566 true/false items that are all stated in the first person (e.g., "I like mechanics magazines."). The 566 items are organized into scales on the MMPI using an empirical keying approach. During its development, the MMPI was administered to both normal individuals and psychiatric patients. An item was assigned to a scale if it separated a clinical group from the normals, regardless of the content of the item. All ten "clinical" scales were created using this approach, and the names of the scales represent the diagnostic groups that these scales were intended to predict. The other three standard scales on the MMPI are validity scales that measure the degree to which a patient may be falsely exaggerating, denying, or otherwise modifying his or her symptoms.

MMPI results are interpreted by referring to patient profiles. Figure 1 shows the MMPI results for one of the 90 patients used in the example data set. Scores are plotted on the profile for each scale, with 50 representing the score of normals, and 70 indicating a significant deviation from normality. Profiles are distinguished primarily on the basis of "peaks," or the scales that have the highest scores. For this patient, the high scores ranked in order are Pa, Sc, F, Pt, Si, and Ma, and this profile is fairly typical of a patient who is diagnosed as having paranoid schizophrenia.

A Few Cautions about Cluster Analysis

Before proceeding with chapters that discuss the basic methodological steps of cluster analysis, a few precautionary generalizations about cluster analysis must be made.

(1) *Most cluster analysis methods are relatively simple procedures that in most cases, are not supported by an extensive body of statistical reasoning.* In other words, most cluster analysis methods are heuristics (simple "rules of thumb"). They are little more than plausible algorithms that can be used to create clusters of cases. This stands in sharp contrast to factor analysis, for instance, which is based upon an extensive body of statistical reasoning. Although many clustering algorithms have important mathematical properties that have been explored in some detail (see Jardin and Sibson, 1971), it is important to recognize the fundamental simplicity of these methods. In doing so, the user is far less likely to make the mistake of reifying the cluster solution.

(2) *Cluster analysis methods have evolved from many disciplines and are inbred with the biases of these disciplines.* This is important to note, because each discipline has its own biases and preferences as to the kinds of questions asked of the data, the types of data thought to be useful in building a classification, and the structure of classifications thought to

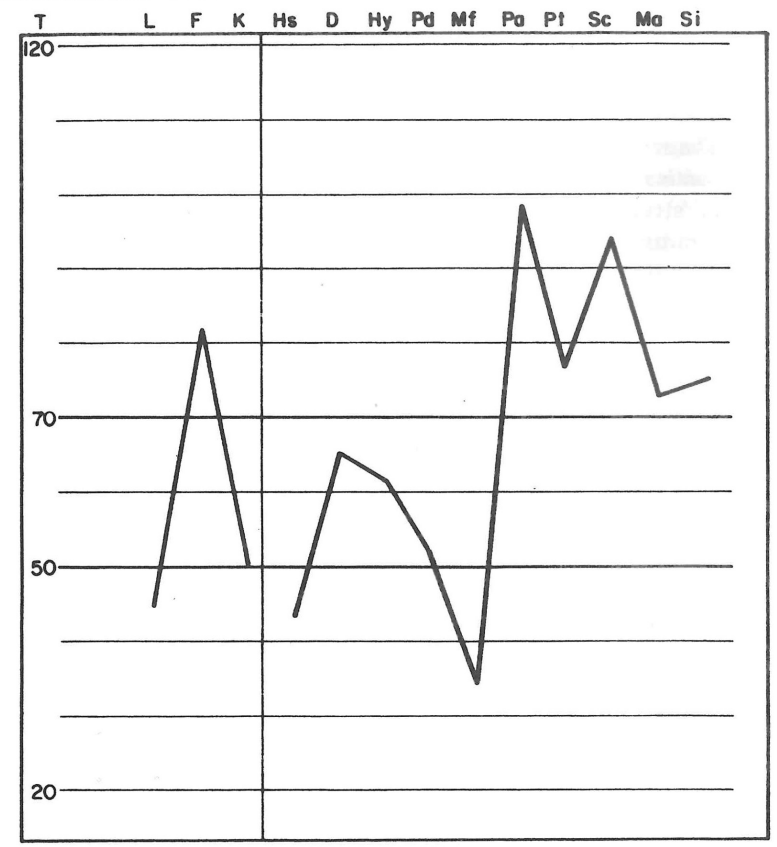

Figure 1: Example of MMPI Profile

be useful. What may be useful in psychology may not be useful to the biologist, and since clustering methods are often no more than plausible rules for creating groups, users must be aware of the biases that often accompany the presentation and description of a clustering method.

(3) *Different clustering methods can and do generate different solutions to the same data set.* The finding that different methods generate different solutions is common in most applied research. One reason for differences in solutions is that clustering methods have evolved from disparate sources that have emphasized different rules of group formation. While this may be a natural outgrowth of disciplinary specialization, it is nevertheless a source of considerable confusion for both novices and sophisticated users of cluster analysis. What is

obviously needed are techniques that can be used to determine what clustering method has discovered the most "natural" groups in a data set. A number of validation procedures have been developed to provide some relief for this problem.

(4) *The strategy of cluster analysis is structure-seeking although its operation is structure-imposing.* That is, clustering methods are used to discover structure in data that is not readily apparent by visual inspection or by appeal to other authority. This strategy differs from that embodied by discriminant analysis, which is more properly described as an identification procedure. This method assigns objects to already existing groups and does not create new ones. Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing. A clustering method will always place objects into groups, and these groups may be radically different in composition when different clustering methods are used. The key to using cluster analysis is knowing when these groups are "real" and not merely imposed on the data by the method.

2. SIMILARITY MEASURES

Terminology

Many terms have been created to describe the important features concerned with the estimation of similarity. As we shall show later (Chapter 5), the development of jargon about cluster analysis within scientific disciplines is an expression of the rapid growth and spread of cluster analysis. Each discipline organizes its terminology in ways that may not necessarily overlap the terminology of other disciplines even if the terms are used to describe the same things. Unless the potential user of cluster analysis is aware of these terminological differences, considerable confusion can result.

The terms "case," "entity," "object," "pattern," and "OTU" (operational taxonomic unit) denote the "thing" being classified; whereas "variable," "attribute," "character," and "feature" denote those aspects of the "things" used to assess their similarity. Another set of important terms are "Q analysis" and "R analysis"; the former refers to the relationships between variables. Cluster analysis, for instance, has been traditionally described as a "Q-mode" technique, whereas factor analysis has been traditionally described as an "R-mode" method.

The potential user of cluster analysis should also note that data matrices are often organized in different ways. In the social sciences the convention is to describe the data set as a matrix consisting of N cases (the rows of the matrix) measured on P variables (the columns of the matrix). In the biological sciences this ordering is reversed, resulting in a $P \times N$ matrix of data. In this book, we reserve the term "raw data" to describe the original $N \times P$ matrix of cases and their variables before the calculation of similarity takes place. Accordingly, we will use the terms "similarity matrix" or "proximity matrix" to describe the $N \times N$ matrix of similarities of the cases after the raw data have been submitted to some measure of similarity.

Even the term "similarity" is not immune from varied meanings, and its synonyms are "resemblance," "proximity," and "association." However, other authors restrict the use of "similarity" to describe a limited set of coefficients. For instance, Everitt (1980) uses the term "similarity coefficient" to denote those measures which Sneath and Sokal (1973) call "association coefficients." Clifford and Stephenson (1975), to confuse things further, restrict the use of the term "association coefficient" to a meaning that is a subset of the definitions by either Everitt or Sneath and Sokal. We use the term "similarity coefficient" (or measure) to describe any type of similarity measure, and adhere to the classification of similarity coefficients proposed by Sneath and Sokal (1973), who subdivide these coefficients into four groups:

(1) correlation coefficients,

(2) distance measures,

(3) association coefficients, and

(4) probabilistic similarity measures.

Later in the chapter each of these groups will be briefly described.

The Concept of Similarity

That things are recognized as similar or dissimilar is fundamental to the process of classification. Despite its apparent simplicity, the concept of similarity, and especially the procedures used to measure similarity, are far from simple. Indeed, the concept of similarity raises basic epistemological problems such as, How can we form useful, abstract concepts to organize what we know? To answer that question, of course, one must be able to categorize things, and the process of categorization requires the lumping together of things that are perceived as similar. The problem of similarity, however, does not lie with the simple recognition that things are either alike or not alike, but instead in the ways in which these concepts are expressed and implemented in scientific research. To be successful, science must be based upon objective, replicable proce-

dures; therefore, the development of statistical procedures to measure more "objectively" the similarity of things is a natural consequence of the necessity for replicable and reliable classifications.

The quantitative estimation of similarity has been dominated by the concept of *metrics*; this approach to similarity represents cases as points in a coordinate space such that the observed similarities and dissimilarities of the points correspond to metric distances between them (Tversky, 1977). The dimensionality of the space is determined by the number of variables used to describe the cases. There are four standard criteria that can be used to judge whether a similarity measure is a true metric. These are

(1) *Symmetry.* Given two entities, x and y, the distance, d, between them satisfies the expression

$$d(x,y) = d(y,x) \geq 0$$

(2) *Triangle inequality.* Given three entities, x, y, z, the distances between them satisfies the expression

$$d(x,y) \leq d(x,z) + d(y,z)$$

Obviously this simply states that the length of any side of a triangle is equal to or less than the sum of the other two sides. This concept has also been called the metric inequality.

(3) *Distinguishability of nonidenticals.* Given two entities x and y,

$$\text{if } d(x,y) \neq 0, \text{ then } x \neq y$$

(4) *Indistinguishability of identicals.* For two identical elements, x and x′,

$$d(x,x') = 0$$

That is, the distance between the two entities is zero.

These are important mathematical properties, and many researchers, most notably Jardine and Sibson (1971) and Clifford and Stephenson (1975), have presented arguments against the routine use of similarity coefficients that do not meet the qualifications of a metric. Many, but not all, distance measures discussed below are metrics. A number of correlation measures are not metric. Coefficients that are not metrics may not be jointly monotonic; that is, the values of different coefficients

used with the same data will not necessarily vary conjointly, raising the disturbing issue that these coefficients could suggest quite different relationships among the entities. The practical significance of determining whether a measure meets the criteria of being metric can be shown by noting that a popular similarity measure, the Pearson product-moment correlation coefficient, certainly fails to meet the third criterion, and as Clifford and Stephenson (1975) suggest, it may well not meet the second (i.e., the triangle inequality) in many applications.

Despite their obvious importance, metrics are by no means the only way to represent the similarity between objects. Certainly on philosophical grounds that are beginning to be supported by psychological research, it is possible to conceive of similarity as the comparison of features; thus the estimation of similarity can be based upon the process of feature matching (Tversky, 1977). This concept of similarity has no inherent dimensionality for the representation of resemblance. Moreover, there is a considerable body of social research in which the similarity between entities is directly estimated. This similarity, for example, may be based on the degree of relationship that exists between entities, and in this type of research, asymmetric similarity values are common. That is, entity A may stand in certain relation to B, but B may not have that degree of relation to A (e.g., Adam may be in love with Betty, but Betty may not like Adam at all). This type of relationship is also common in economics, where one nation can import more goods from another nation than it exports to that nation. Asymmetry presents special problems in the calculation of similarity coefficients. Tversky (1977) provides a good introduction to this issue.

The potential user of cluster analysis should be aware that many types of similarity exist, and that while many of the coefficients and measures commonly used in quantitative approaches to classification are metrics, there are alternatives to the use of these measures that may be appropriate and necessary within the context of research. The choice of similarity measure, then, should be embedded ultimately within the design of research, which is itself determined by the theoretical, practical, and philosophical context of the classification problem.

The Choice of Variables

Before describing popular coefficients used in the calculation of similarity, a brief digression needs to be made about the choice of variables and about data transformations prior to the calculation of similarity. The choice of variables to be used with cluster analysis is one of the most critical steps in the research process, but, unfortunately, it is one of the least understood as well. The basic problem is to find that set

of variables that best represents the concept of similarity under which the study operates. Ideally, variables should be chosen within the context of an explicitly stated theory that is used to support the classification. The theory is the basis for the rational choice of the variables to be used in the study. In practice, however, the theory that supports classification research is often implicit, and in this situation it is difficult to assess the relevance of the variables to the problem.

The importance of using theory to guide the choice of variables should not be underestimated. The temptation to succumb to a naive empiricism in the use of cluster analysis is very strong, since the technique is ostensibly designed to produce "objective" groupings of entities. By "naive empiricism" we mean the collection and subsequent analysis of as many variables as possible in hope that the "structure" will emerge if only enough data are obtained. While empirical studies are important to the progress of any science, those that adopt an implicit naive empiricist perspective are dangerous in the context of cluster analysis because of the heuristic nature of the technique and the many unsolved problems that have plagued its application (Everitt, 1979).

In most statistical analyses the data are routinely standardized by some appropriate method. If the normality of a variable is in question, a logarithmic or other transformation is often performed. If the data are not of the same scale values, they are commonly standardized to a mean of 0 and to unit variance. There is some controversy, however, as to whether standardization should be a routine procedure in cluster analysis. As Everitt (1980) notes, standardization to unit variance and mean of 0 can reduce the differences between groups on those variables that may well be the best discriminators of group differences. It would be far more appropriate to standardize variables *within* groups (i.e., within clusters), but obviously this cannot be done until the cases have been placed into groups.

Edelbrock (1979) has noted that variables in multivariate data sets may have different distribution parameters across groups; thus standardization may not constitute an equivalent transformation of these variables and could possibly change the relationships between them. However, his Monte Carlo studies of the effects of standardization on subsequent analyses using the correlation coefficient and various hierarchical clustering methods did not reveal substantial differences between the use of standardized versus nonstandardized variables in the resulting classifications. Milligan (1980) has also found that standardization appears to have only a minor effect on the results of a cluster analysis. Others, most notably Matthews (1979), have shown that standardization did have a negative effect on the adequacy of the results of a

cluster analysis when compared to an "optimal" classification of the cases under study.

The situation regarding standardization is far from clear. Users with substantially different units of measurement will undoubtedly want to standardize them, especially if a similarity measure such as Euclidean distance is to be used. The decision to standardize should be made on a problem-to-problem basis, and users should be aware that results may differ solely on the basis of this factor, although the magnitude of the effect will vary from data set to data set.

Other types of data transformation are possible, and many of these have been used concurrently with cluster analysis. Factor analysis or principal components analysis is often used when the researcher knows that the variables used in the study are highly correlated. The uncritical use of highly correlated variables to compute a measure of similarity is essentially an implicit weighting of these variables. That is, if three highly correlated variables are used, the effect is the same as using only one variable that has a weight three times greater than any other variable. Principal components analysis and factor analysis can be used to reduce the dimensionality of the data, thereby creating new, uncorrelated variables that can be used as raw data for the calculation of similarity between cases. Once again, there is controversy surrounding this procedure. Factor analysis tends to blur the relationship between clusters because it assumes that factor scores are normally distributed. The effect of factor analysis is to transform the data in such a way that any modes present are merged, resulting in variables that are normally distributed. Rohlf (1970) has noted that principal components analysis tends to maintain the representation of widely separated clusters in a reduced space but also minimizes—and thus blurs—the distances between clusters or groups that are not widely separated.

The problem of whether or not to weight variables has also aroused considerable controversy. Most of this debate has taken place within the biological sciences. Weighting is simply the manipulation of a value of a variable such that it plays a greater or lesser role in the measurement of similarity between two cases (Williams, 1971). While the concept of weighting is simple, its practice is difficult, and very few guidelines exist. Williams describes five types of weighting, the most common being the a priori manipulation of variables. Sneath and Sokal (1973) argue strongly against a priori weighting, and suggest that the appropriate way to measure similarity is to give all variables equal weight. This advice, however, must be tempered with the understanding that the Sneath and Sokal view of clustering is considered a radically empirical approach to the creation of classifications. In many instances it may well make sense

to weight a particular variable a priori if there are good theoretical reasons for this and there are well-defined procedures under which weighting can occur. While the issue of weighting has not yet become an issue of debate in the social sciences, researchers using clustering methods should be aware of the controversy.

Similarity Measures

Now that the problems of variable selection and data transformation have been discussed, a presentation of popular similarity coefficients can be offered. As noted earlier, there are four types of similarity measures: (1) correlation coefficients, (2) distance measures, (3) association coefficients, and (4) probabilistic similarity coefficients. Each of these methods has advantages and disadvantages that must be considered before a decision is made to use one. Although all four types have been used extensively by numerical taxonomists and others in the biological sciences, only correlation and distance coefficients have had widespread use in the social sciences. Correspondingly, we devote more discussion to these two types of measures.

CORRELATION COEFFICIENTS

These coefficients, often called angular measures because of their geometric interpretation, are among the most frequently used measures of similarity in the social sciences. The most popular is the product-moment correlation coefficient suggested by Karl Pearson. Originally defined for use as a method to correlate variables, it has been used in quantitative classification to determine the correlation between cases. In this context, the coefficient is defined as

$$r_{jk} = \frac{\Sigma(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\Sigma(x_{ij} - \bar{x}_j)^2 (x_{ik} - \bar{x}_k)^2}}$$

where x_{ij} is the value of variable i for case j and $\bar{x}_j$ is the mean of all values of the variable for case j. The method is used with ratio or interval scale variables, and in the case of binary data it is transformed into the familiar phi coefficient. The value of the coefficient ranges from -1 to +1, and a value of zero indicates no relationship between the cases. Since the mean of each case is summed across all variables of each case, standard significance tests of r have no obvious meaning.

The correlation coefficient is frequently described as a *shape* measurement, in that it is insensitive to differences in the magnitude of the variables used to compute the coefficient. As Williams (1971) notes, Pearson's r is sensitive to shape because of its implicit standardization of each case across all variables. This property is of special importance to disciplines such as psychology, sociology, and anthropology that often describe data in terms of profiles. While a profile is formally defined as nothing more than a vector of attribute values for a case (Sneath and Sokal, 1973) the comparison of each case, as represented by a vector but displayed graphically as a profile, is often the desired end product of a classification. For instance, the MMPI data used in this book are often plotted to create profiles of different individuals (see Figure 1).

One of the major drawbacks of the use of the correlation coefficient as a similarity measure is its sensitivity to shape at the expense of the magnitude of differences between the variables. As first demonstrated by Cronbach and Gleser (1953), the similarity between profiles can be decomposed into three parts: *shape*, the pattern of dips and rises across the variables; *scatter*, the dispersion of the scores around their average; and *elevation* (level or size), the mean score of the case over all of the variables. That the product-moment correlation coefficient is sensitive only to shape means that two profiles can have a correlation of +1.0 and yet not be identical (i.e., the profiles of each case do not pass through the same points). Figure 2 shows two MMPI profiles, one with a solid line and the other with a dashed line. Their shapes are identical. Although the correlation between these two profiles is +1.0, they are not truly identical because one is elevated. Thus, a high correlation can occur between profiles as long as the measurements of one profile are in a linear relationship to another. Some information is lost, therefore, when the correlation coefficient is used, and it is possible that misleading results can be obtained if the effects of dispersion and elevation on profile data are not also considered.

There are other potential limitations of the coefficient. It often fails to satisfy the triangle inequality, and as many have pointed out, the use of the method to calculate the correlation of cases does not make statistical sense, because one must obtain the mean value across different variable types rather than across cases, as in the standard use of the method. The meaning of the "mean" across these variables is far from clear.

Despite these drawbacks, the coefficient has been used successfully in a wide variety of research applications involving cluster analysis. Hamer and Cunningham (1981) have demonstrated that the correlation coefficient is superior in its ability to reduce the total number of misclassifications when used with a consistent clustering method as compared to

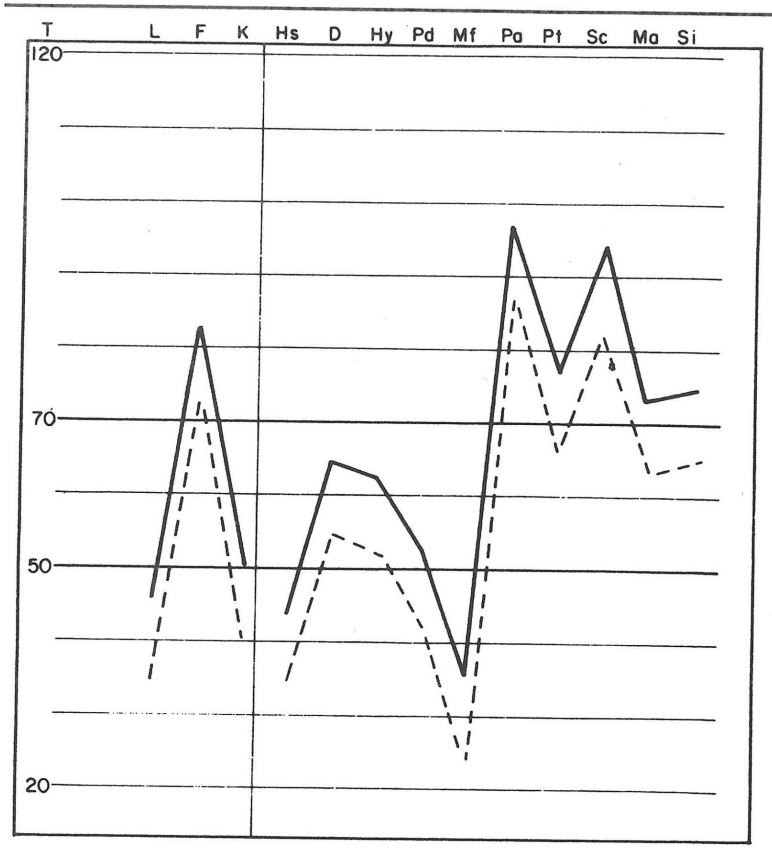

Figure 2: MMPI Profiles (r = 1.0)

other similarity coefficients. Paradoxically, correlation proves to be of value precisely because it is not affected by dispersion and size differences between the variables. Of considerable importance to the success of this study, however, was that the researchers were able to postulate their need for a shape coefficient because they believe that dispersion and size effects on the profile data were created by differences in the experience and judgment of raters of job classifications, and were thus not due to the intrinsic variability of the job classifications themselves.

DISTANCE MEASURES

Because of their intuitive appeal, distance measures have enjoyed widespread popularity. Technically, they are best described as *dissimi-*

larity measures; most of the more popular coefficients demonstrate similarity by high values within their ranges, but distance measures are scaled in the reverse. Two cases are identical if each one is described by variables with the same magnitudes. In this case, the distance between them is zero. Distance measures normally have no upper bounds, and are scale-dependent. Among the more popular representations of distance is Euclidean distance, defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

where d_{ij} is the distance between cases i and j, and x_{ik} is the value of the k^{th} variable for the i^{th} case. To avoid the use of the square root, the value of distance is often squared, and this is usually indicated by the term d_{ij}^2. As might be expected, this expression is referred to as "squared Euclidean distance."

Other types of distance can be defined, and another popular measure is the Manhattan distance, or city-block metric, which is defined as

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$

Other metrics can be defined, but most are specific forms of the special class of metric distance functions known as Minkowski metrics, defined in a general form as

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}$$

There are other distances that are not Minkowski metrics, and the most important of these is Mahalanobis D^2, also called generalized distance (Mahalanobis, 1936). This metric is defined as

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

where Σ is the pooled within-groups variance-covariance matrix, and X_i and X_j are vectors of the values of the variables for cases i and j. Unlike Euclidean or other Minkowski metrics, this metric incorporates the

correlations among variables by the inclusion of the variance-covariance matrix. When the correlation between variables is zero, Mahalanobis D^2 is equivalent to squared Euclidean distance.

Despite their importance, Euclidean and other distance metrics suffer from serious problems, among the most critical of which is that the estimation of the similarity between cases is strongly affected by elevation differences. Variables with both large size differences and standard deviations can essentially swamp the effects of other variables with smaller absolute sizes and standard deviations. Moreover, distance metrics are also affected by transformations of the scale of measurement of the variables, in that Euclidean distance will not preserve distance rankings (Everitt, 1980). In order to reduce the effect of the relative size of the variables, researchers routinely standardize the variables to unit variance and means of zero before the calculation of distance. As noted above, this type of data transformation may lead to other kinds of problems.

Skinner (1978) has proposed a useful and feasible way in which to use both correlation and Euclidean distance to calculate the similarity of profile data so that it is possible to determine which of the factors (shape, size, and dispersion) have contributed to the estimation of similarity. The basic strategy is similar to that proposed by Guertin (1966), in which correlation is used to create homogeneous groups based on shape, and each shape group is then divided with a distance measure into subgroups with similar size and dispersion characteristics (Skinner, 1978). Skinner's procedure, however, represents a considerable advance over this method in that it develops a composite similarity function that integrates both distance and correlation in a computational strategy that tends to minimize measurement error in estimation of profile similarity.

Since the need to standardize data occurs frequently in applied data analysis, a brief example showing the effects of standardization on correlation and distance may be helpful. The data used to illustrate the problem are four MMPI profiles. Each of these profiles represents a psychotic patient with severe psychopathology.

The initial similarity measure to be used with these four profiles is the Pearson product-moment correlation coefficient. The results are shown below:

	A	B	C	D
A	xxxx	.776	.702	.742
B	(3)	xxxx	.729	.779
C	(6)	(5)	xxxx	.936
D	(4)	(2)	(1)	xxxx

The values of the correlations are shown in the upper triangular half of the matrix. These values show that all four profiles have very similar shapes, but, in particular, profiles C and D are close to being identical (r_{CD} = .936). Shown in the lower triangular portion of the matrix are the rank orderings of the similarity values from the most similar (1) to least similar (6). The significance of the rank orderings will become apparent below.

When Euclidean distances are calculated, the following matrix is obtained:

	A	B	C	D
A	xxxx	266	732	736
B	(2)	xxxx	532	465
C	(5)	(4)	xxxx	144
D	(6)	(3)	(1)	xxxx

Notice how different the scaling of the distance coefficients are when compared to the correlation coefficients. Remember, the values of the distance coefficients have no absolute meanings. Again, patients C and D appear to be the most similar with d_{CD} = 144, but there is no clear indication of exactly how good a value of 144 is. In general, the pattern of similarities appears about the same when using correlation and distance, but differences do exist. In particular, the least similar patients using correlation as the similarity measure were patients A and C (r_{AC} = .702). However, Euclidean distance suggests that patients A and D are the least similar (d_{AD} = 736).

To confuse the picture even more, suppose we decide to standardize the data. (Standardization was actually performed on the basis of statistics from the entire 90 case data set.) If the product-moment correlation is used to assess the similarity of the four profiles after standardization, the similarity matrix is as displayed below:

	A	B	C	D
A	xxxx	.602	.284	.433
B	(2)	xxxx	.367	.584
C	(6)	(5)	xxxx	.804
D	(4)	(3)	(1)	xxxx

Notice how different the values of the correlations appear to be when comparing standardized versus nonstandardized data. With nonstandardized data, r_{AC} = .702, but with standardized data, r_{AC} = .284. In both

instances, r_{AC} is the smallest value in the matrix, but on standardized data, the value of the correlation coefficient suggests that patients A and C are not at all similar while with nonstandardized data, the absolute value of the correlation (r = .706) suggests that A and C in fact are reasonably similar.

Finally, the dissimilarity matrix below shows the Euclidean distances among the patients using standardized data.

	A	B	C	D
A	xxxx	.704	2.572	2.071
B	(1)	xxxx	2.141	1.304
C	(6)	(5)	xxxx	.870
D	(4)	(3)	(2)	xxxx

Again the values change as a function of standardization. However, since the magnitudes of the value of a Euclidean distance coefficient has no inherent meaning, this change is not particularly important. What is important is the relative change. The most dramatic change is that the Euclidean distance coefficient for the standardized data shows patients A and B to be the most similar pair, while the other three similarity matrices showed patients C and D to be the most similar.

To conclude this brief comparison, it is important to note that all four matrices yielded separate, nonidentical rankings of the similarity values. This point is important because it demonstrates what a dramatic effect the choice of a similarity coefficient and a data transformation can have on the relationships in the resulting similarity matrix.

ASSOCIATION COEFFICIENTS

This type of measure is used to establish similarity between cases described by binary variables. It is easiest to discuss these coefficients by reference to the familiar 2 × 2 association table which 1 refers to the presence of a variable and 0 to its absence.

	1	0
1	a	b
0	c	d

A large number (> 30) of these coefficients have been proposed, and it is unrealistic to attempt to describe comprehensively the full range of these measures. As might be expected, most of these coefficients were first defined in biological systematics, although it is likely that some of

the simplest of them have been reinvented in a number of disciplines. Few of the measures have been extensively tested, and many have been dropped because of questionable features. Good references for those interested in these coefficients are Sneath and Sokal (1973), Clifford and Stephenson (1975), and Everitt (1980). There are three measures, however, that have been used extensively and deserve special consideration. These are the simple matching coefficient, Jaccard's coefficient, and Gower's coefficient.

The simple matching coefficient is defined as

$$S = \frac{(a + d)}{(a + b + c + d)}$$

where S is the similarity between the two cases which ranges from 0 to 1. As Sneath and Sokal (1973) note, this coefficient cannot be easily transformed into a metric. However, considerable effort has been devoted to the establishment of approximate confidence limits, one of the few methods of this type so honored (Goodall, 1967). The coefficient takes into account the joint absence of a variable (as indicated in the d cell of the association, matrix).

Jaccard's coefficient, defined as

$$S = a/(a + b + c),$$

avoids the use of joint absences of a variable in the calculation of similarity (it omits the d cell from consideration). Like the simple matching coefficient, it ranges from 0 to 1. It has seen extensive use in the biological sciences as a result of the debate over the inclusion of so-called negative matches (joint absence of a variable). Biologists have noted that if the simple matching coefficient is used, some cases would appear very similar primarily because they both lacked the same features rather than because the features they did have were shared. In contrast, Jaccard's coefficient is concerned only with features that have positive co-occurrences.

The problem of whether to include negative matches has not apparently been an issue in most social sciences, but the problem has arisen in archaeology. If an object is not found with a burial, its absence may be due to either cultural prescriptions or natural processes of disintegration and attrition. It would be inappropriate to base the estimation of similarity between two burials upon the joint absence of an artifact if it is impossible to know which of the two possible explanations is responsible for its absence.

To provide a brief example comparing the simple matching coefficient and Jaccard's coefficient, six data points from the burial data will be examined.

1	C	M	N	1	0	0	1	0	0	0	0
5	C	F	E	0	0	1	0	0	0	1	0
8	T	M	N	0	1	0	1	1	0	0	0
14	T	F	E	1	0	0	0	1	0	1	0
18	A	M	E	1	1	0	1	1	0	1	1
24	A	F	E	1	0	0	0	1	1	1	0

Consider cases 1 (a male child of nonelite status; i.e., CMN for child, male, nonelite) and 8 (male adolescent of nonelite status, i.e., TMN for teenage, male, nonelite). The 2 × 2 association matrix of common features between these two cases is

		TMN	
		1	0
CMN	1	1	1
	0	2	4

That is, these two cases have only one shared artifact. However, four artifacts were absent from both burials. Thus,

$$S = .625 \ (= 5/8)$$

However,

$$J = .250 \ (= 1/4)$$

In other words, while the simple matching coefficient suggests that cases CMN and TMN are reasonably similar, Jaccard's coefficient implies they are not. The entire 6 × 6 similarity matrix using the simple matching coefficient is

	CMN	CFE	TMN	TFE	AME	AFE
CMN	–	.500	.625	.625	.500	.500
CFE		–	.375	.625	.250	.500
TMN			–	.500	.625	.375
TFE				–	.625	.875
AME					–	.500
AFE						–

For Jaccard's coefficient, the similarity matrix is

	CMN	CFE	TMN	TFE	AME	AFE
CMN	–	.000	.250	.250	.333	.200
CFE		–	.000	.250	.143	.200
TMN			–	.200	.500	.166
TFE				–	.500	.750
AME					–	.429
AFE						–

These matrices are reasonably similar. For instance, both matrices suggest that cases TFE, AME, and AFE—the older elite burials— are the most similar. However, differences do exist. The two children, cases CMN and CFE, are totally dissimilar according to Jaccard's coefficient, but appear to be relatively similar according to the simple matching coefficient.

Another feature of these matrices is the number of "ties." With the simple matching coefficient there are five pairs of cases for which S = .625, and six cases where S = .500. In fact, in the fifteen cells of the 6 × 6 similarity matrix, there are only five (!) unique values of S. As we shall show later, some clustering methods perform poorly when so many ties are present in the similarity matrix.

Gower's coefficient is unique because it permits the simultaneous use of variables of different scales of measurement in the estimation of similarity. Proposed by Gower (1971), it is defined as

$$s_{ij} = \sum_{k=1}^{p} S_{ijk} \Big/ \sum_{k=1}^{p} W_{ijk}$$

where W_{ijk} is a weighting variable valued at 1 if a comparison of variable k is considered valid and 0 if it is not. S_{ijk} is a similarity "score" based upon the outcome of the comparison of variable k across cases i and j. In the case of binary variables, W_{ijk} is zero when variable k is not known for one or both individuals under comparison (Everitt, 1980). In the case of so-called negative matches, W_{ijk} is also set to zero. It should be clear that if the data are all binary, the coefficient is identical to Jaccard's.

To demonstrate how this coefficient works, the burial data set has been modified to include two new variable types: stature (measured in

centimeters, thus a quantitative variable) and estimated energy expenditure in grave construction or excavation (measured on an ordinal scale with ranks 1, 2, and 3, or low, moderate, and high, respectively). Four cases have been modified:

1	C	M	N	1	0	0	1	0	0	0	0	69	1
7	T	M	N	1	1	0	1	0	0	0	0	167	2
18	A	F	E	1	1	0	1	1	0	1	1	179	3
25	A	M	E	1	0	0	0	1	1	1	1	158	3

For binary data, S_{ijk} is calculated according to the following scoring system:

case i		1	1	0	0
case j		1	0	1	0
score	S_{ijk}	1	0	0	0
weight	W_{ijk}	1	1	1	0

For ordinal data, $S_{ijk} = 1$ when the values of the comparison are identical, and 0 when they are not. Finally, for quantitative data, the equation

$$S_{ijk} = 1 - |x_{ik} - x_{jk}|/R_k$$

where x_{ik} is the score of case i on variable k and R_k is the range of variable k. The resulting similarity matrix from these manipulations for the four cases is

	CMN	TMN	AME	AFE
CMN	–	.527	.285	.170
TMN	–	–	.554	.239
AME	–	–	–	.726
AFE	–	–	–	–

The coefficient has a number of appealing features beyond its ability to accommodate mixed data types. These include its metric qualities and its flexibility, in that the method can be easily modified to include negative matches in the estimation of similarity by simply modifying the binary weighting system. That the coefficient has seen relatively little use in the social sciences can probably be attributed to its failure to appear in any major clustering software packages (see Chapter 5).

PROBABILISTIC SIMILARITY COEFFICIENTS

Coefficients of this type are radically different from those described above in that, technically, the similarity between two cases is not actually calculated. Instead, this type of measure works directly upon the raw data. When forming clusters, the information gain (in the Shannon sense) of the combination of two cases is evaluated, and that combination of cases that provides the least information gain is fused. Another important point about probabilistic measures is that they can be used only with binary data. No workable schemes for using this type of measure with quantitative and qualitative variables has yet been developed. These coefficients have not yet appeared in the social sciences, but they have been used extensively by numerical taxonomists and ecologists for at least a decade. Comprehensive summaries of these measures can be found in Sneath and Sokal (1973) and Clifford and Stephenson (1975).

Suggested Readings

The most valuable and detailed discussion of similarity coefficients relevant to cluster analysis can be found in Sneath and Sokal (1973). These authors devote 74 pages to the discussion of similarity and provide formulae for most of the measures they discuss. Clifford and Stephenson (1975) also provide a useful discussion of similarity measures of relevance to cluster analysis.

More broadly, the theoretical issues associated with similarity are discussed in Hartigan (1967) and Tversky (1977). Skinner's (1978) discussion of shape, elevation, and scatter are very relevant to many uses of similarity measures in social science research. The last three references are important because the concept of similarity is crucial to the formation of clusters. Clusters, after all, are usually defined as groups of *similar* entities. Although most discussions of cluster analysis emphasize the procedures for creating clusters, the choice of a measure of similarity is crucial in any clustering study.

3. A REVIEW OF CLUSTERING METHODS

On the Nature of Clusters

The primary reason for the use of cluster analysis is to find groups of similar entities in a sample of data. These groups are conveniently referred to as clusters. There is no standard or even useful definition of