In L. G. Grimm and P. R. Yarnold (Eds.) (2000). Reading and understanding more multivariate statistics. Washington, DC: American Psychological Association.

Cluster Analysis

Joseph F. Hair, Jr., and William C. Black

Cluster analysis is a group of multivariate techniques whose primary purpose is to assemble objects based on the characteristics that they possess. Cluster analysis classifies *objects* (e.g., respondents, products, or other entities), so that each object is similar to others in the cluster with respect to a predetermined selection criterion. The resulting clusters of objects should then exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. Thus, if the classification is successful, the objects within the clusters will be close together when plotted geometrically and different clusters will be farther apart. A common use of cluster analysis in clinical psychology is the identification of types (clusters) of disorders. For example, a researcher may want to know if it is important to identify different types of attention deficit hyperactivity disorder (ADHD). A cluster analysis may reveal that a syndrome that specifies attention deficits should, perhaps, be separated from a syndrome that emphasizes hyperactivity. The cluster analysis can reveal what symptoms discriminate the two categories. Subsequent validation, as described later in this chapter, is needed to justify the use of the proposed ADHD clusters.

In cluster analysis, the concept of the variate is again a central issue but in a different way from other multivariate techniques. The *cluster variate* is the set of variables representing the characteristics used to compare objects in the cluster analysis. Because the cluster variate in-

This chapter is from *Multivariate Data Analysis*, by J. R. Hair, Jr., R. E. Anderson, R. L. Tatham, and W. C. Black, 1998, pp. 469–518.

cludes only the variable used to compare objects, it determines the "character" of the objects. Cluster analysis is the only multivariate technique that does not estimate the variate empirically but uses the variate as specified by the researcher. The focus of cluster analysis is on the comparison of objects based on the variate, not on the estimation of the variate itself. This makes the researcher's definition of the variate a critical step in cluster analysis.

Cluster analysis has been referred to as "Q analysis," "typology," "classification analysis," and "numerical taxonomy." This variety of names is due in part to the use of clustering methods in such diverse disciplines as psychology, biology, sociology, economics, engineering, and business. Although the names differ across disciplines, the methods all have a common dimension: classification according to natural relationships (Aldenderfer & Blashfield, 1984; Anderberg, 1973; Bailey, 1994; Green & Carroll, 1978; Punj & Stewart, 1983; Sneath & Sokal, 1973). This common dimension represents the essence of all clustering approaches. As such, the primary value of cluster analysis lies in the classification of data, as suggested by "natural" groupings of the data themselves. Cluster analysis is comparable with factor analysis in its objective of assessing structure; cluster analysis differs from factor analysis in that cluster analysis groups objects, whereas factor analysis is primarily concerned with grouping variables.

Cluster analysis is useful in many situations. For example, a researcher who has collected data with a questionnaire may be faced with a large number of observations that are meaningless unless classified into manageable groups. Using cluster analysis, the researcher can perform data reduction objectively by decreasing the information from an entire population or sample to information about specific, smaller subgroups. For example, if we can understand the attitudes of a population by identifying the major groups within the population, then we have reduced the data for the entire population into profiles of a number of groups. In this fashion, the researcher has a more concise, understandable description of the observations, with minimal loss of information.

Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine already stated hypotheses. For example, a researcher may believe that attitudes toward the consumption of diet versus regular soft drinks could be used to separate soft drink consumers into logical segments or groups. Cluster analysis can classify soft drink consumers by their attitudes about

diet versus regular so
profiled for demogra
clinical psychologist c
to identify potential s

These examples
of cluster analysis. Ra
for grouping all living
based on personality
yses of marketers, cl
grouping individuals.
in almost every area
on the use of cluster
ing of the technique

Yet with the ben
analysis can be chara
ential. Cluster analysi
inferences from a sa
exploratory techniqu
membership for any
ments of the proced
varying one or more
clusters, regardless o
Finally, the cluster s
as the basis for the s
evant variables can
Thus, the researche
cision involved in p

Brief Descripti

Before continuing w
tions of published s

A Typology of Custo

Singh (1990) perfo
system of styles of c
sions of complaint
the type of respons

diet versus regular soft drinks, and the resulting clusters, if any, can be profiled for demographic similarities and differences. Alternatively, a clinical psychologist could use cluster analysis to form symptom clusters to identify potential subtypes of a disorder.

These examples are just a small fraction of the types of applications of cluster analysis. Ranging from the derivation of taxonomies in biology for grouping all living organisms to psychological classifications that are based on personality and other personal traits, to the segmentation analyses of marketers, cluster analysis has always had a strong tradition of grouping individuals. The result has been an explosion of applications in almost every area of inquiry, creating not only a wealth of knowledge on the use of cluster analysis but also the need for a better understanding of the technique to minimize its misuse.

Yet with the benefits of cluster analysis come some caveats. Cluster analysis can be characterized as descriptive, atheoretical, and noninferential. Cluster analysis has no statistical basis on which to draw statistical inferences from a sample to a population and is primarily used as an exploratory technique. The solutions are not unique because the cluster membership for any number of solutions is dependent on many elements of the procedure; many different solutions can be obtained by varying one or more elements. Moreover, cluster analysis always creates clusters, regardless of the "true" existence of any structure in the data. Finally, the cluster solution is totally dependent on the variables used as the basis for the similarity measure. The addition or deletion of relevant variables can have a substantial effect on the resulting solution. Thus, the researcher must take care in assessing the effect of each decision involved in performing a cluster analysis.

Brief Descriptions

Before continuing with the details of cluster analysis, two brief descriptions of published studies that use cluster analysis are provided.

A Typology of Customer Complaints

Singh (1990) performed a cluster analysis to develop a categorization system of styles of consumer complaint behavior (CCB). Three dimensions of complaint intentions or behaviors that differ on the basis of the type of response (actions directed at the seller, negative word-of-

mouth, or complaints to third parties) were recorded from a random sample of store customers seen in a 2-year period. Based on these three behaviors, cluster analysis was used to identify groups of similar individuals. Four clusters of consumer groups were identified: (a) no action; (b) voice actions only; (c) voice and private actions; and (d) voice, private, and third-party actions. The results of testing whether the response styles would reproduce differences in actual behavior were offered as support for the validity of the cluster solution. Finally, a number of demographic, personality–attitudinal, and situational variables—identified in the research literature as important to understanding consumer complaints—were used to profile the CCB styles. In addition, the author used multiple discriminant analysis to determine the relative importance of demographic variables for each cluster. This study extends previous research and demonstrates the multifaceted nature of complaint styles. Such findings should be of interest to retail managers by increasing knowledge about customers and improving the handling of customer complaints.

Police Officers' Perceptions of Rape

Campbell and Johnson (1997) examined how police officers defined rape, irrespective of the legal definition in their state. Officers' narrative responses to the following question were content analyzed:

> As you know, it's the legislators that make the laws and decide how to define crimes and what punishments will be. But you are actually in the community, dealing with victims and criminals. Based on your work as a police officer, how do you define rape–sexual assault?

Three clusters of definitions were retained. The smallest cluster (19% of officers) was characterized as "force definition of rape" and corresponded closely to the state's legal definition of rape. In contrast to the most salient feature of rape (force), as defined by the state in which the study was conducted, a second cluster was comprised of 31% of the sample and emphasized penetration and consent ("consent definition of rape"). The third cluster included 50% of the sample and was labeled "mixed definition of rape." Officers provided definitions that departed from the state's legal definition and included such stereotypic attitudes as "Sometimes a guy can't stop himself; he gets egged on by the girl."

Campbell and Johnson (1997) then provided validity data by comparing the groups on measures not included in the cluster analysis. Included among the variables that significantly differed across groups

were job title, training in sexual assault law, attitudes toward women, and attitudes toward interpersonal violence. Because only one cluster represented officer definitions that closely corresponded with state law and this was the smallest cluster, the authors recommended that law enforcement agencies pay closer attention to formal training in the legal definition of rape and sexual assault.

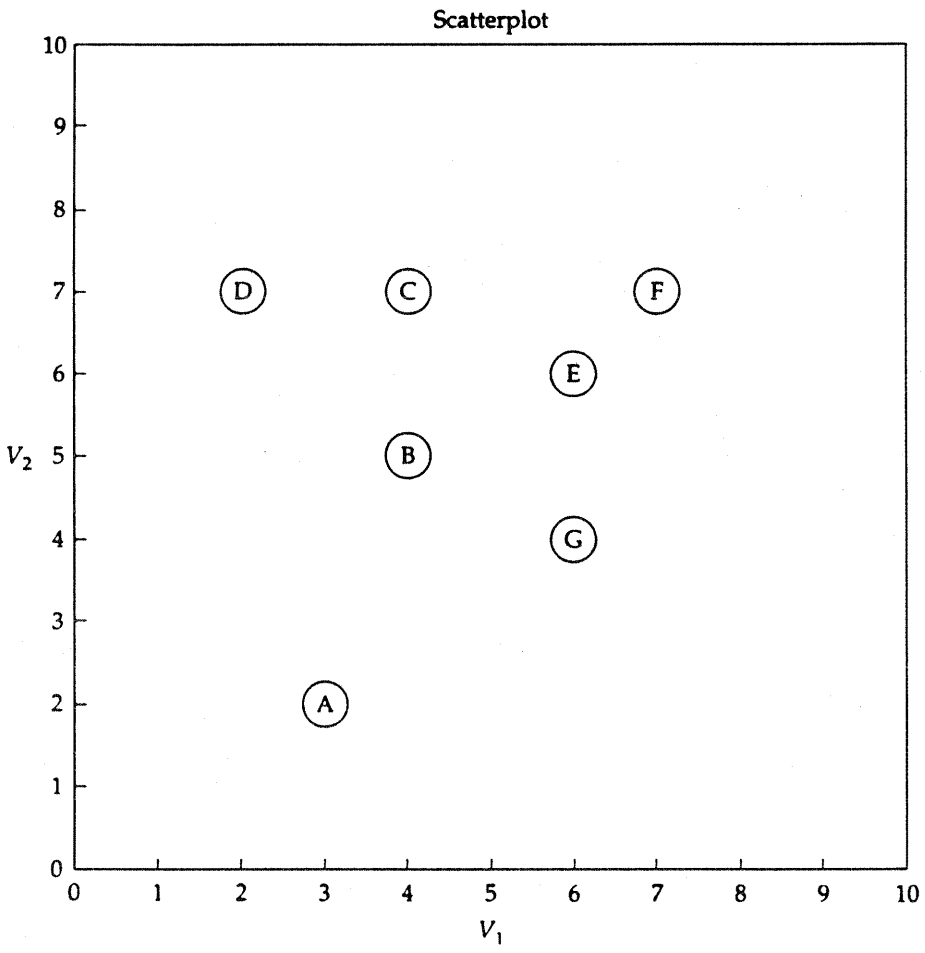
How Does Cluster Analysis Work?

The nature of cluster analysis can be illustrated by a simple bivariate example. Suppose a marketing researcher wishes to determine market segments (clusters) in a small community on the basis of patterns of loyalty to brands and stores. A small sample of seven of the respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty, V_1 (store loyalty) and V_2 (brand loyalty), were measured for each respondent on a scale of 0 to 10 (0 = *not loyal at all*; 10 = *highly loyal*). The values for each respondent are shown in Figure 5.1, along with a scatter diagram depicting each observation on the variables.

The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups. To accomplish this task, however, the research must address three questions. First, how does one measure similarity? It requires a method for simultaneously comparing observations on the two clustering variables (V_1 and V_2).

Several methods are possible, including the correlation between objects; a measure of association used in other multivariate techniques; or perhaps a measure of their proximity in two-dimensional space, such that the distance between observations indicates similarity. Second, how does one form clusters? No matter how similarity is measured, the procedure must group those observations that are most similar into a cluster. This procedure must determine for each observation which other observations it will be grouped with. Third, how many groups does one form? Any number of "rules" might be used, but the fundamental task is to assess the "average" similarity across clusters, such that as the average increases, the clusters become less similar. The researcher then faces a trade-off: fewer clusters versus less homogeneity. Simple structure, in striving toward parsimony, is reflected in as few clusters as possible. Yet as the number of clusters decreases, the homogeneity within the clusters necessarily decreases. Thus, a balance must be found be-

Figure 5.1



Data values and scatterplot of seven observations based on two variables.

tween defining the most basic structure (fewer clusters) that still achieves the necessary level of similarity within the clusters. Once we have procedures for addressing each issue, we can perform a cluster analysis.

Measuring Similarity

We illustrate a cluster analysis for the seven observations (respondents) using simple procedures for each of the issues. First, similarity is measured as the Euclidean distance (straight line) between each pair of

Table 5.1

Proximity Matri

Observation
A
B
C
D
E
F
G

observations. Tabl
respondent. In us
remember that sn
Observations E an
the most dissimila

Forming Clusters

We must next dev
in this chapter, m
poses here, we use
est) observations
clusters. We apply
its own "cluster,"
tions are in a singl
it moves in a step
tions. It is also an
the combination

Table 5.2 det
ing the initial state
Then clusters are
cluster remains. S
and combines the
Next, Step 2 find
three pairs have t
and C). Let us st
cluster, E was com

Table 5.1

Proximity Matrix of Euclidean Distances Between Observations

	Observation						
Observation	A	B	C	D	E	F	G
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

observations. Table 5.1 contains measures of proximity between each respondent. In using distance as the measure of proximity, we must remember that smaller distances indicate greater similarity, such that Observations E and F are the most similar (1.414), whereas A and F are the most dissimilar (6.403).

Forming Clusters

We must next develop a procedure for forming clusters. As we see later in this chapter, many methods have been proposed; but for our purposes here, we use this simple rule: identify the two most similar (closest) observations not already in the same cluster, and combine their clusters. We apply this rule repeatedly, starting with each observation in its own "cluster," and combine two clusters at a time until all observations are in a single cluster. This is termed a *hierarchical procedure* because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an *agglomerative method* because clusters are formed by the combination of existing clusters.

Table 5.2 details the steps of the hierarchical process, first depicting the initial state with all seven observations in single-member clusters. Then clusters are joined in the agglomerative process until only one cluster remains. Step 1 identifies the two closest observations (E and F) and combines them into a cluster, moving from seven to six clusters. Next, Step 2 finds the next closest pairs of observations. In this case, three pairs have the same distance of 2.000 (E and G, C and D, and B and C). Let us start with E and G. Although G is a single member cluster, E was combined in the prior step with F, so the cluster formed

Table 5.2

Agglomerative Hierarchical Clustering Process

Agglomeration process			Cluster solution		
Step	Minimum distance between unclustered observations[a]	Observation pair	Cluster membership	No. of clusters	Overall similarity measure (average within-cluster distance)
Initial solution			(A) (B) (C) (D) (E) (F) (G)	7	0
1	1.414	E–F	(A) (B) (C) (D) (E–F) (G)	6	1.414
2	2.000	E–G	(A) (B) (C) (D) (E–F–G)	5	2.192
3	2.000	C–D	(A) (B) (C–D) (E–F–G)	4	2.144
4	2.000	B–C	(A) (B–C–D) (E–F–G)	3	2.234
5	2.236	B–E	(A) (B–C–D–E–F–G)	2	2.896
6	3.162	A–B	(A–B–C–D–E–F–G)	1	3.420

[a]Euclidean distance between observations.

at this stage now
single member c
the two-member
there are Cluster
and G).

The next sm
(E and B, B and
however, as each
two existing clust
two three-membe
step (6) is to con
observations) intc
are distances sma
because they are

The hierarch
several ways. Figu
process is hierarcl
of nested groupin
represent the pro
variables in the sc
mon approach is
cess in a treelike
tion coefficient, i
This approach is
It also depicts the
unwieldy when th

Determining the N

A hierarchical me
case ranging from
should we choose?
ters, homogeneity
homogeneous pos
ture with seven cl
tion for its descrip
of the clusters.

In this examp
the average distan
right-most column

at this stage now has three members: G, E, and F. Step 3 combines the single member clusters of C and D, whereas Step 4 combines B with the two-member cluster C–D that was formed in Step 3. At this point, there are Cluster 1 (A), Cluster 2 (B, C, and D), and Cluster 3 (E, F, and G).

The next smallest distance is 2.236 for three pairs of observations (E and B, B and G, C and E). We use only one of these distances, however, as each observation pair contains a member from each of the two existing clusters (B, C, D vs. E, F, G). Thus, Step 5 combines the two three-member clusters into a single six-member cluster. The final step (6) is to combine Observation A with the remaining cluster (six observations) into a single cluster at a distance of 3.162. Note that there are distances smaller than or equal to 3.162, but they are not used because they are between members of the same cluster.

The hierarchical clustering process can be portrayed graphically in several ways. Figure 5.2 illustrates two such methods. First, because the process is hierarchical, the clustering process can be shown as a series of nested groupings (see Figure 5.2A). This process, however, can only represent the proximity of the observations for two or three clustering variables in the scatter plot or three-dimensional graph. A more common approach is the *dendrogram*, which represents the clustering process in a treelike graph. The horizontal axis represents the agglomeration coefficient, in this instance the distance used in joining clusters. This approach is useful in identifying outliers, such as Observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.

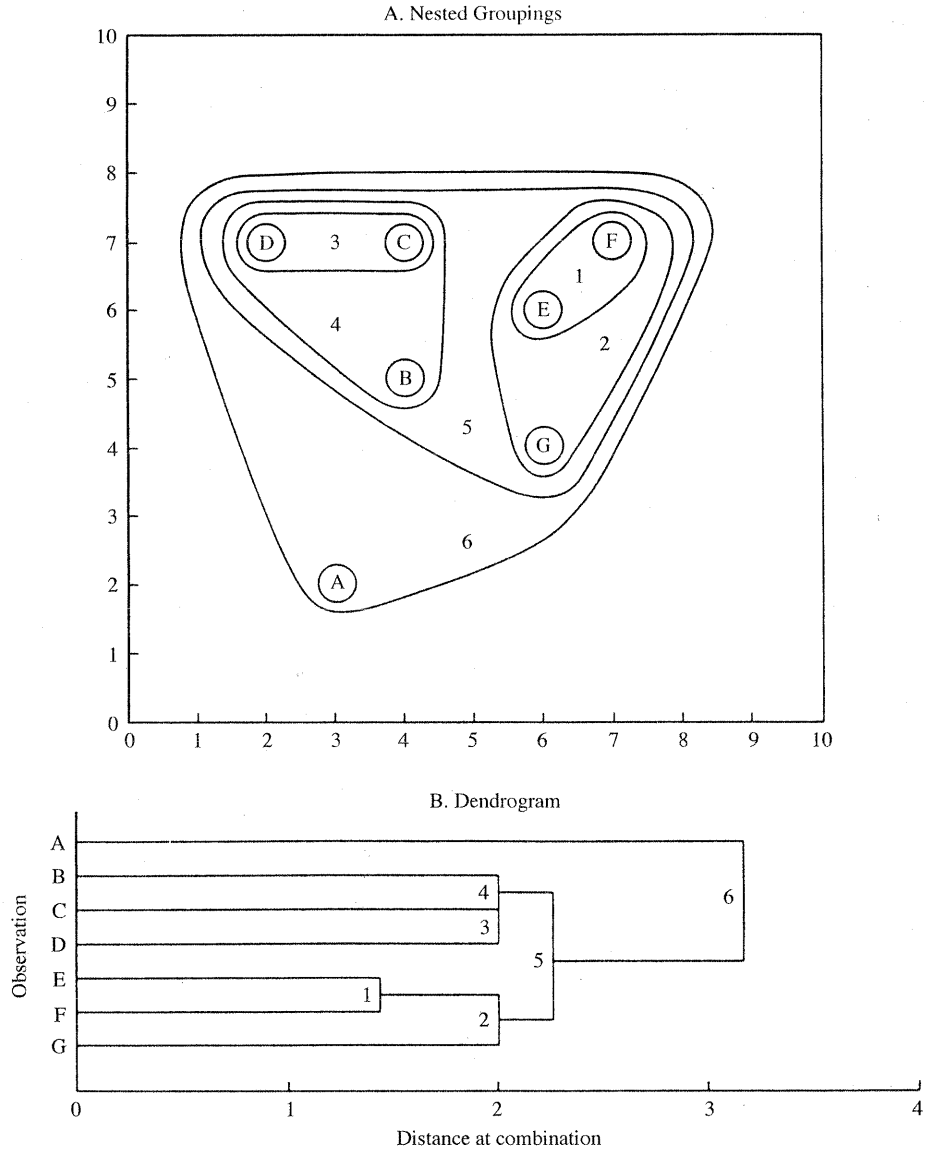
Determining the Number of Clusters in the Final Solution

A hierarchical method results in a number of cluster solutions, in this case ranging from a one-cluster to a six-cluster solution. But which one should we choose? We know that as we move from single-member clusters, homogeneity decreases. So why not stay at seven clusters, the most homogeneous possible? The problem is that we did not define any structure with seven clusters. So the researcher must view each cluster solution for its description of structure balanced against the homogeneity of the clusters.

In this example, we use a simple measure of homogeneity, namely, the average distances of all observations within clusters (refer to the right-most column of Table 5.2). In the initial solution with seven clus-

"Euclidean distance between observations.

Figure 5.2



Graphical portrayals of the hierarchical clustering process.

ters, our overall similarity measure is 0—no observation is paired with another. For the six-cluster solution, the overall similarity is the distance between the two observations (1.414) joined at Step 1. Step 2 forms a three-member cluster (E, F, and G), so that the overall similarity mea-

sure is the mea
(2.000), and F a
two-member clu
overall average
clusters this ma
where the avera

Now, how
cluster solution?
possible that stil
overall similarit
increases in the
similar. In our e
two observation
member cluster
measure does n
ing other cluste
clusters. But wh
member cluster

This indica
cluster that was
cluster solution
can also see that
indicating that e
the last step, wh
geneity marked
vation A as com
member of the
independent of
of cluster soluti
appropriate for
and the single c

As one wo
cluster solution
sidered by mar
methods have b
tions, it still falls
to accept as the
bivariate case be
and social scien
are measured o

sure is the mean of the distances between E and F (1.414), E and G (2.000), and F and G (3.162), for an average of 2.192. In Step 3, a new two-member cluster is formed with a distance of 2.000, which causes the overall average to fall slightly to 2.144. We can proceed to form new clusters this manner until a single-cluster solution is formed (Step 6), where the average of all distances in the distance matrix is 3.412.

Now, how do we use this overall measure of similarity to select a cluster solution? Remember that we are trying for the simplest structure possible that still represents homogeneous groupings. If we monitor the overall similarity measure as the number of clusters decreases, large increases in the overall measure indicates that two clusters were not that similar. In our example, the overall measure increases when we first join two observations (Step 1) and then again when we make our first three-member cluster (Step 2). But in the next two steps (3 and 4), the overall measure does not change substantially. This indicates that we are forming other clusters with essentially the same homogeneity of the existing clusters. But when we get to Step 5, which combines the two three-member clusters, we see a large increase.

This indicates that joining these two clusters resulted in a single cluster that was markedly less homogeneous. We would consider the cluster solution of Step 4 much better than that found in Step 5. We can also see that in Step 6 the overall measure actually decreases slightly, indicating that even though the last observation remained separate until the last step, when it was joined it did not change the cluster homogeneity markedly. However, given the rather unique profile of Observation A as compared with the others, it might best be designated as a member of the *entropy group*, those observations that are outliers and independent of the existing clusters. Thus, when reviewing the range of cluster solutions, the three-cluster solution of Step 4 seems the most appropriate for a final-cluster solution, with two equally sized clusters and the single outlying observation.

As one would probably realize by now, the selection of the final-cluster solution requires substantial researcher judgment, and it is considered by many as too subjective. Even though more sophisticated methods have been developed to assist in evaluating the cluster solutions, it still falls to the researcher to decide as to the number of clusters to accept as the final solution. Cluster analysis is rather simple in this bivariate case because the data are two dimensional. In most marketing and social sciences research studies, however, more than two variables are measured on each object, and the situation is much more complex

with many more observations. We discuss how the researcher can use more sophisticated procedures to deal with the increased complexity of "real-world" applications in the remainder of this chapter.

Cluster Analysis Decision Process

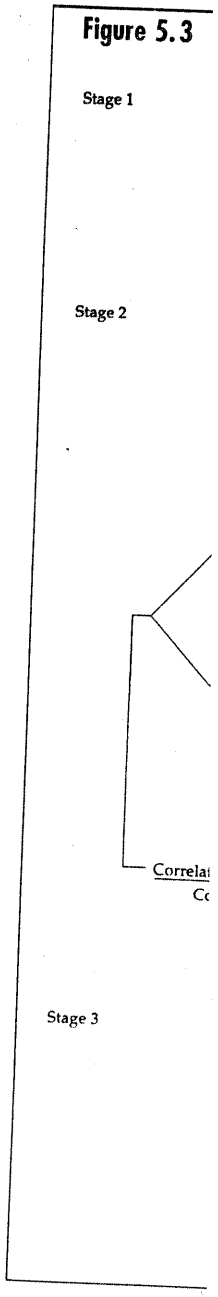
Cluster analysis can be viewed from a six-stage model-building approach. Starting with research objectives that can be either exploratory or confirmatory, the design of a cluster analysis deals with the partitioning of the data set to form clusters, the interpretation of the clusters, and the validation of the results. The partitioning process determines how clusters may be developed. The interpretation process involves understanding the characteristics of each cluster and developing a name or label that appropriately defines its nature. The third process involves assessing the validity of the cluster solution (i.e., determining its stability and generalizability), along with describing the characteristics of each cluster to explain how they may differ on relevant dimensions such as demographics. The following sections detail all these issues through the six stages of the model-building process. Figure 5.3 depicts the first three stages of the cluster analysis decision tree.

Stage 1: Objectives

The primary goal of cluster analysis is to partition a set of objects into two or more groups based on the similarity of the objects for a set of specified characteristics (cluster variate). In forming homogeneous groups, the researcher can achieve any of three objectives:

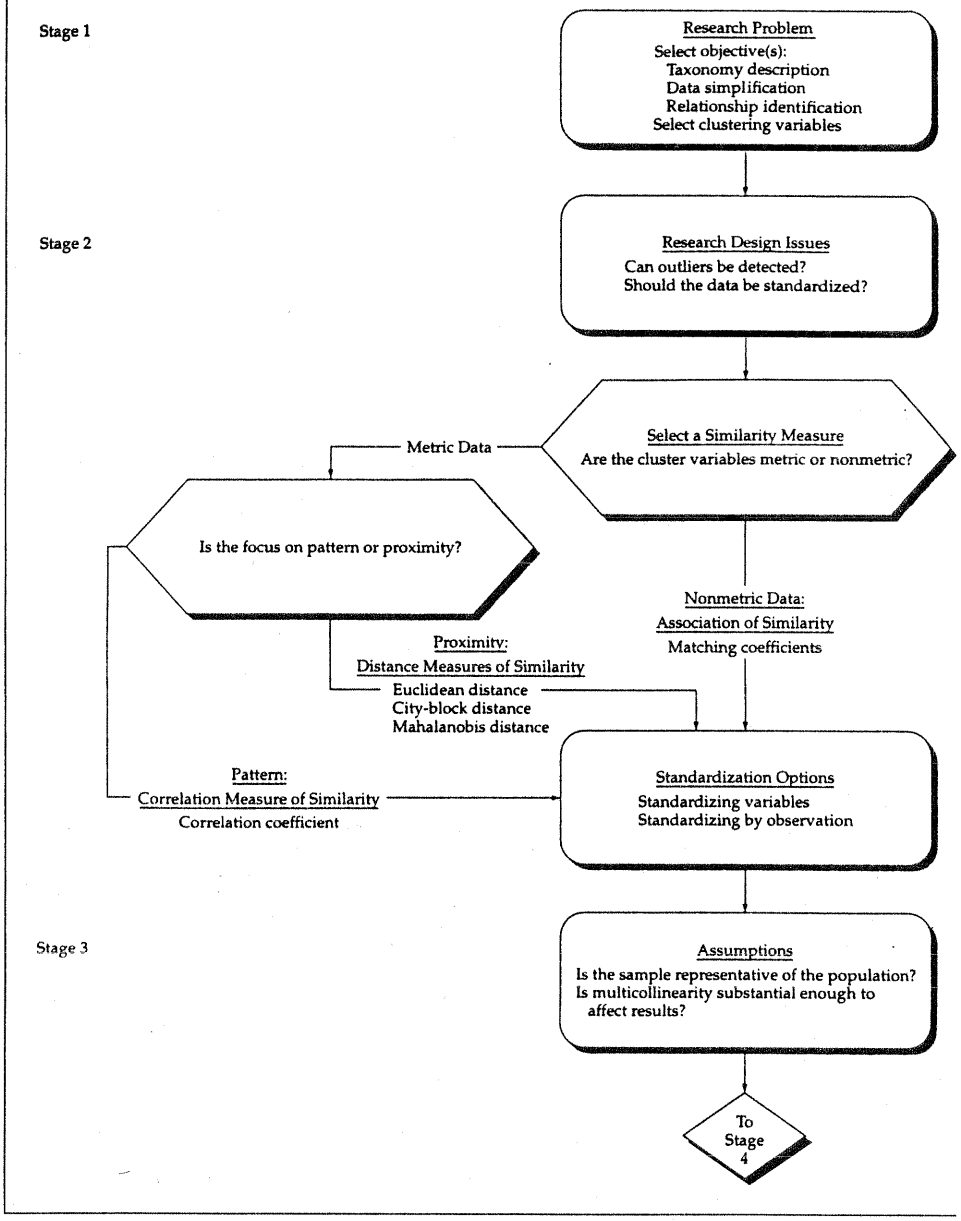
1. *Taxonomy description.* The most traditional use of cluster analysis has been for exploratory purposes and the formation of a *taxonomy*—an empirically based classification of objects. As described earlier, cluster analysis has been used in a range of applications for its partitioning ability. However, cluster analysis can also generate hypotheses related to the structure of the objects. Although viewed principally as an exploratory technique, cluster analysis can be used for confirmatory purposes. If a proposed structure can be defined for a set of objects, cluster analysis can be applied and a proposed *typology* (theoretically based classification) can be compared with that derived from the cluster analysis.

Figure 5.3



Stages 1–3 of th

Figure 5.3



Stages 1–3 of the cluster analysis decision diagram.

2. *Data simplification.* In the course of deriving a taxonomy, cluster analysis also achieves a simplified perspective on the observations. With a defined structure, the observations can be grouped for further analysis. Whereas factor analysis attempts to provide "dimensions" or structure to variables, cluster analysis performs the same task for observations. Thus, instead of viewing all of the observations as unique, they can be viewed as members of a cluster and profiled by its general characteristics.
3. *Relationship identification.* With the clusters defined and the underlying structure of the data represented in the clusters, the researcher has a means of revealing relationships among the observations that was perhaps not possible with the individual observations. For example, one could use discriminant analysis to discover which variables, not used in the cluster analysis, discriminate among the cluster groups.

In any application, the objectives of cluster analysis cannot be separated from the selection of variables used to characterize the objects to be clustered. Whether the objective is exploratory or confirmatory, the researcher effectively limits the possible results by the variables selected for use. The derived clusters can only reflect the inherent structure of the data as defined by the variables.

Selecting the variables to be included in the cluster variate must be done with regard to both theoretical–conceptual and practical considerations. Any application of cluster analysis must have some rationale on which variables are selected. Whether the rationale is based on an explicit theory, past research, or supposition, the researcher must realize the importance of including only those variables that (a) characterize the objects being clustered and (b) relate specifically to the objectives of the cluster analysis. The cluster analysis technique has no means of differentiating the relevant from irrelevant variables. It only derives the most consistent, yet distinct, groups of objects across all variables. The inclusion of an irrelevant variable increases the chance that outliers will be created on these variables, which can have a substantive effect on the results. Thus, one should never include variables indiscriminately, but instead should carefully choose the variables with the research objective as the criterion for selection.

In a practical vein, cluster analysis can be dramatically affected by the inclusion of only one or two inappropriate or undifferentiated variables (Milligan, 1980). The researcher is always encouraged to examine the results and to eliminate the variables that are not distinctive (i.e.,

that do not diff
dure allows the
on only those v

Stage 2: Resear

With the objecti
address three qu
outliers be detec
object similarity
Many approache
has been evalua
these questions,
different results
factor analysis, i
our discussion r
amples of the m
the practical lim

The import
stages becomes a
is seeking struct
through a selecte
the possible part
partitioning 25 c
10^{15} possible par
sions of the rese
solutions as "cor
and the choice o
effect than perha

Irrelevant Variable

In its search for
of irrelevant varia
thus, the use of
grade the reliabil
accordingly, inhil
validity compone
liers (objects that
resent either (a)
of the general po
the population th

that do not differ significantly) across the derived clusters. This procedure allows the cluster techniques to maximally define clusters based on only those variables exhibiting differences across the objects.

Stage 2: Research Design

With the objectives defined and variables selected, the researcher must address three questions before starting the partitioning process: (a) Can outliers be detected and, if so, should they be deleted? (b) How should object similarity be measured? and (c) Should the data be standardized? Many approaches can be used to answer these questions. However, none has been evaluated sufficiently to provide a definitive answer to any of these questions, and unfortunately, many of the approaches provide different results for the same data set. Thus, cluster analysis, along with factor analysis, is much more of an art than a science. For this reason, our discussion reviews these issues in a general way by providing examples of the most commonly used approaches and an assessment of the practical limitations where possible.

The importance of these issues and the decisions made in later stages becomes apparent when we realize that although cluster analysis is seeking structure in the data, it must actually impose a structure through a selected methodology. Cluster analysis cannot evaluate all of the possible partitions because for even the relatively small problem of partitioning 25 objects into 5 nonoverlapping clusters there are 2.4×10^{15} possible partitions (Anderberg, 1973). Instead, based on the decisions of the researcher, the technique identifies one of the possible solutions as "correct." From this viewpoint, the research design issues and the choice of methodologies made by the researcher have greater effect than perhaps with any other multivariate technique.

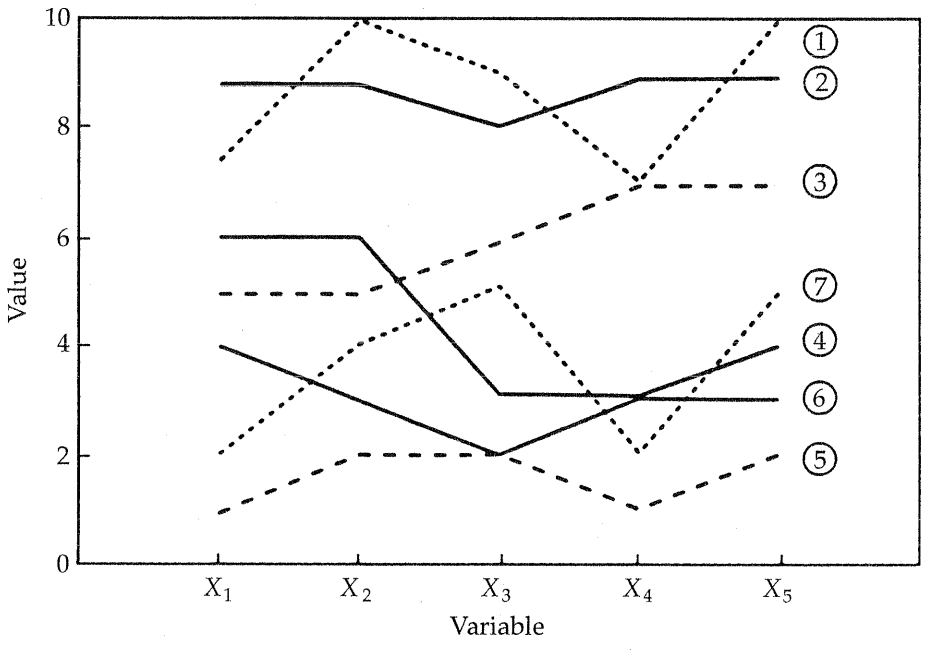
Irrelevant Variables and Outliers

In its search for structure, cluster analysis is sensitive to the inclusion of irrelevant variables. Such variables essentially constitute "noise," and thus, the use of irrelevant variables to define typologies serves to degrade the reliability (or reproducibility) of the group assignments and, accordingly, inhibits the performance of group status assignment in the validity component of the study. Cluster analysis is also sensitive to outliers (objects that are very different from all others). Outliers can represent either (a) truly aberrant observations that are not representative of the general population or (b) an undersampling of actual groups in the population that causes an underrepresentation of the groups in the

sample. In both cases, the outliers distort the true structure and make the derived clusters unrepresentative. For this reason, a preliminary screening for outliers is always necessary. Probably the easiest way to conduct this screening is to prepare a graphic profile diagram, such as that shown in Figure 5.4. The *profile diagram* lists the variables along the horizontal axis and the variable values along the vertical axis. Each point on the graph represents the value of the corresponding variable, and the points are connected to facilitate visual interpretation. Profiles for all objects are then plotted on the graph, a line for each object. Outliers are those objects with very different profiles, most often characterized by extreme values on one or more variables.

Obviously, such a procedure becomes cumbersome with a large number of objects (observations) or variables. For the observations shown in Figure 5.4, there is no obvious outlier that has all extremely high or low values. Just as in detecting multivariate outliers in other multivariate techniques, however, outliers may also be defined as having unique profiles that distinguish them from all of the other observations. Also, they may emerge in the calculation of similarity. By whatever

Figure 5.4



Profile diagram for screening outliers (1–7).

means used, c
their represent
if deemed unr
however, the re
from the samp

Similarity Measu

The concept o
similarity is a m
jects to be clus
between variab
comparable pr
defining simila
bined into a sir
we used correl
any object can
measure. The c
objects togethe

Interobjec
three methods
lational measu
Each method
dent on both i
and distance m
association me

Correlation

probably come
pair of objects
relating two se
so that the col
variables. Thus
numbers is the
two objects. Hi
denote a lack c
type factor ana

Correlatio
of patterns acr
examining the
5.4. A correlati

means used, observations identified as outliers must be assessed for their representativeness of the population and deleted from the analysis if deemed unrepresentative. As in other instances of outlier detection, however, the researcher should exhibit caution in deleting observations from the sample, as they may distort the actual structure of the data.

Similarity Measures

The concept of similarity is fundamental to cluster analysis. *Interobject similarity* is a measure of correspondence or resemblance between objects to be clustered. In factor analysis, we create a correlation matrix between variables that is then used to group variables into factors. A comparable process occurs in cluster analysis. Here, the characteristics defining similarity are first specified. Then, the characteristics are combined into a similarity measure calculated for all pairs of objects, just as we used correlations between variables in factor analysis. In this way, any object can be compared with any other object through the similarity measure. The cluster analysis procedure then proceeds to group similar objects together into clusters.

Interobject similarity can be measured in a variety of ways, but three methods dominate the applications of cluster analysis: (a) correlational measures, (b) distance measures, and (c) association measures. Each method represents a particular perspective on similarity, dependent on both its objectives and the type of data. Both the correlational and distance measures require metric (continuous) data, whereas the association measures are for nonmetric (categorical) data.

Correlational measures. The interobject measure of similarity that probably comes to mind first is the correlation coefficient between a pair of objects measured on several variables. In effect, instead of correlating two sets of variables, we invert the objects' X variables matrix so that the columns represent the objects and the rows represent the variables. Thus, the correlation coefficient between the two columns of numbers is the correlation (or similarity) between the profiles of the two objects. High correlations indicate similarity, and low correlations denote a lack of it. This procedure is followed in the application of Q-type factor analysis (see chapter 11).

Correlational measures represent similarity by the correspondence of patterns across the characteristics (X variables). This is illustrated by examining the example of seven observations (cases) shown in Figure 5.4. A correlational measure of similarity looks not at the magnitude of

the values but instead at the patterns of the values. In Table 5.3, which contains the correlations among these seven observations, we can see two distinct groups. First, Cases 1, 5, and 7 all have similar patterns and corresponding high positive intercorrelations (i.e., for Cases 1 and 5, $r = .963$; Cases 1 and 7, $r = .891$; and Cases 5 and 7, $r = .963$). Likewise, Cases 2, 4, and 6 also have high positive correlations among themselves but low or negative correlations with the other observations. Case 3 has low or negative correlations with all other cases, thereby perhaps forming a group by itself. Thus, correlations represent patterns across the variables much more than the magnitudes. Correlational measures, however, are rarely used because the emphasis in most applications of cluster analysis is on the magnitudes of the objects, not the patterns of values.

Distance measures. Although correlational measures have an intuitive appeal and are used in many other multivariate techniques, they are not the most commonly used measure of similarity in cluster analysis. Distance measures of similarity, which represent similarity as the proximity of observations to one another across the variables in the cluster variate, are the similarity measure most often used. Distance measures are actually a measure of dissimilarity, with larger values denoting lesser similarity. Distance is converted into a similarity measure by using an inverse relationship. A simple illustration of this was shown in the hypothetical example in which clusters of observations were defined based on the proximity of observations to one another when each observation's scores on two variables were plotted graphically (see Figure 5.2).

Comparison to correlational measures. The difference between correlational and distance measures can be seen by referring again to Figure 5.4. Distance measures focus on the magnitude of the values and portray as similar cases that are close together but that may have very different patterns across the variables. Table 5.3 also contains distance measures of similarity for the seven cases, and we see a very different clustering of cases emerging than that found when using the correlational measures. With smaller distances representing greater similarity, Cases 1 and 2 form one group, while Cases 4, 5, 6, and 7 make up another group. These groups represent those with higher versus lower values. A third group, consisting of only Case 3, can be seen as differing from the other two groups as it has values that are both low and high. Whereas the two clusters using distance measures have different members than those using correlations, Case 3 is unique in either measure

Table 5.3

Calculating C

Case
1
2
3
4
5
6
7

Case	1
1	1.(
2	−0.1
3	0.(
4	0.(
5	0.9
6	−0.4
7	0.8

Case	1
1	nc
2	3.3
3	6.8
4	10.3
5	15.7
6	13.
7	11.3

Note. nc = distar
R. E. Anderson,
Prentice-Hall, Inc
NJ.

able 5.3, which
ns, we can see
ar patterns and
Cases 1 and 5,
963). Likewise,
ong themselves
ons. Case 3 has
perhaps form-
erns across the
onal measures,
applications of
the patterns of

s have an intui-
echniques, they
in cluster anal-
imilarity as the
ariables in the
used. Distance
urger values de-
ilarity measure
this was shown
ations were de-
ther when each
hically (see Fig-

between corre-
ig again to Fig-
the values and
may have very
ontains distance
a very different
ing the correla-
eater similarity,
nd 7 make up
ier versus lower
een as differing
h low and high.
different mem-
either measure

Table 5.3

Calculating Correlational and Distance Measures of Similarity

Original Data

	Variables				
Case	X_1	X_2	X_3	X_4	X_5
1	7	10	9	7	10
2	9	9	8	9	9
3	5	5	6	7	7
4	6	6	3	3	4
5	1	2	2	1	2
6	4	3	2	3	3
7	2	4	5	2	5

Similarity Measure: Correlation

	Case						
Case	1	2	3	4	5	6	7
1	1.000						
2	−0.147	1.000					
3	0.000	0.000	1.000				
4	0.087	0.516	−0.824	1.000			
5	0.963	−0.408	0.000	−0.060	1.000		
6	−0.466	0.791	−0.354	0.699	−0.645	1.000	
7	0.891	−0.516	0.165	−0.239	0.963	−0.699	1.000

Similarity Measure: Euclidean Distance

	Case						
Case	1	2	3	4	5	6	7
1	nc						
2	3.32	nc					
3	6.86	6.63	nc				
4	10.24	10.20	6.00	nc			
5	15.78	16.19	10.10	7.07	nc		
6	13.11	13.00	7.28	3.87	3.87	nc	
7	11.27	12.16	6.32	5.10	4.90	4.36	nc

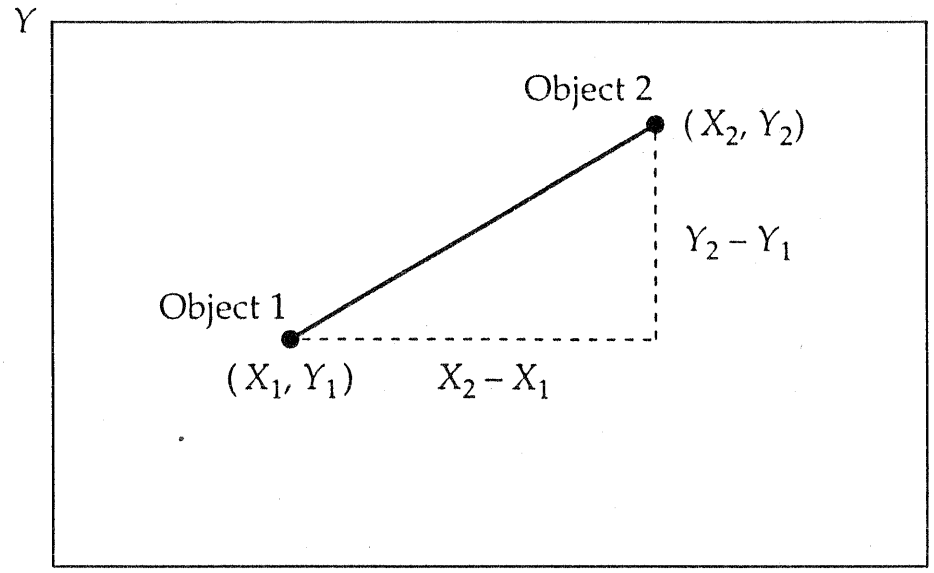
Note. nc = distances not calculated. From *Multivariate Data Analysis*, by J. R. Hair, Jr., R. E. Anderson, R. L. Tatham, and W. C. Black, 1998, p. 485. Copyright 1998 by Prentice-Hall, Inc. Adapted with permission of Prentice-Hall, Inc., Upper Saddle River, NJ.

of similarity. The choice of a correlational measure, rather than the more traditional distance measure, requires a different interpretation of the results by the researcher. Clusters based on correlational measures may not have similar values but may instead have similar patterns. Distance-based clusters have more similar values across the set of variables, but the patterns can be quite different.

Types of distance measures. Several distance measures are available. The most commonly used is Euclidean distance. An example of how Euclidean distance is obtained is shown geometrically in Figure 5.5.

Suppose that two points in two dimensions have coordinates (X_1, Y_1) and (X_2, Y_2), respectively. The Euclidean distance between the points is the length of the hypotenuse of a right triangle, as calculated by the formula under the figure. This concept is easily generalized to more than two variables. The Euclidean distance is used to calculate several specific measures, one being the simple Euclidean distance (calculated as described above) and the other is the squared, or absolute, Euclidean distance, in which the distance value is the sum of the squared differ-

Figure 5.5



An example of Euclidean distance between two objects measured on two variables.

ences without taking the square root. The squared Euclidean distance has the advantage of not taking the square root, which speeds computations markedly, and is the recommended distance measure for the centroid and Ward's methods of clustering (see below).

Several options not based on the Euclidean distance are also available. One of the most widely used alternative measures involves replacing the squared differences by the sum of the absolute differences of the variables. This procedure is the absolute, or city-block, distance function. The *city-block approach* to calculating distances may be appropriate under certain circumstances, but it causes several problems (Shephard, 1966). One such problem is the assumption that the variables are not correlated with one another; if they are correlated, the clusters are not valid. Other measures that use variations of the absolute differences or the powers applied to the differences (other than just squaring the differences) are also available in most cluster programs.

Effect of unstandardized data values. A problem faced by all of the distance measures that use unstandardized data is the inconsistencies between cluster solutions when the scale of the variables is changed. For example, suppose three objects, A, B, and C, are measured on two variables, probability of purchasing brand X (in percentages) and amount of time spent viewing commercials for brand X (in minutes or seconds). The values for each observations are shown in Table 5.4(A).

From this information, distance measures can be calculated. In our example, we calculate three distance measures for each object pair: (a) simple Euclidean distance, (b) squared or absolute Euclidean distance, and (c) city-block distance. First, we calculate the distance values based on purchase probability and the viewing time in minutes. These distances, with smaller values indicating greater proximity and similarity, and their rank order are shown in Table 5.4B. As we can see, the most similar objects (with the smallest distance) are B and C, followed by A and C, with A and B the least similar (or least proximal). This ordering holds for all three distance measures, but the relative similarity or dispersion between objects is the most pronounced in the squared Euclidean distance measure.

The ordering of similarities can change markedly with only a change in the scaling of one of the variables. If we measured the viewing time in seconds instead of minutes, then the rank orders would change (see Table 5.4C). Although B and C are still the most similar, the pair A–B is now next most similar and is almost identical to the similarity of B–C. Yet when we used minutes of viewing time, pair A–B was the

Table 5.4

Variations in Distance Measures Based on Alternative Data Scales

A. Original Data

Object	Purchase probability	Commercial viewing time	
		Minutes	Seconds
A	60	3.0	180
B	65	3.5	210
C	63	4.0	240

B. Distance Measures Based on Minutes

	Simple Euclidean distance		Squared or absolute Euclidean distance		City-block distance	
Object pair	Value	Rank	Value	Rank	Value	Rank
A–B	5.025	3	25.25	3	5.5	3
A–C	3.162	2	10.00	2	4.0	2
B–C	2.062	1	4.25	1	2.5	1

C. Distance Measures Based on Seconds

	Simple Euclidean distance		Squared or absolute Euclidean distance		City-block distance	
Object pair	Value	Rank	Value	Rank	Value	Rank

	distance		Euclidean distance		City-block distance	
Object pair	Value	Rank	Value	Rank	Value	Rank
A–B	5.025	3	25.25	3	5.5	3
A–C	3.162	2	10.00	2	4.0	2
B–C	2.062	1	4.25	1	2.5	1

C. Distance Measures Based on Seconds

	Simple Euclidean distance		Squared or absolute Euclidean distance		City-block distance	
Object pair	Value	Rank	Value	Rank	Value	Rank
A–B	30.41	2	925	2	35	3
A–C	60.07	3	3,609	3	63	2
B–C	30.06	1	904	1	32	1

D. Distance Measures Based on Standardized Values

	Standardized values		Simple Euclidean distance		Squared or absolute Euclidean distance		City-block distance	
Object pair	Purchase probability	Minutes/seconds of viewing time	Value	Rank	Value	Rank	Value	Rank
A–B	−1.06	−1.0	2.22	2	4.95	2	2.99	2
A–C	0.93	0.0	2.33	3	5.42	3	3.19	3
B–C	0.13	1.0	1.28	1	1.63	1	1.79	1

least similar by a substantial margin. What has occurred is that the scale of the viewing time variable has dominated the calculations, making purchase probability less significant in the calculations. The reverse was true, however, when we measured viewing time in minutes because purchase probability was dominant in the calculations. The researcher should thus note the tremendous effect that variable scaling can have on the final solution. Standardization of the clustering variables, whenever possible conceptually, should be used to avoid such instances as found in our example. The issue of standardization is discussed in the following section.

A commonly used measure of Euclidean distance that directly incorporates a standardization procedure is the *Mahalanobis distance.* The Mahalanobis approach not only performs a standardization process on the data by scaling in terms of the standard deviations but also sums the pooled within-group variance–covariance, which adjusts for intercorrelations among the variables. Highly intercorrelated sets of variables in cluster analysis can implicitly overweight one set of variables in the clustering procedures, as is discussed in the next section. In short, the Mahalanobis generalized distance procedure computes a distance measure between objects comparable to the R^2 in regression analysis. Although many situations are appropriate for use of the Mahalanobis distance, many computer software programs do not include it as a measure of similarity. In such cases, the researcher usually selects the squared Euclidean distance.

In attempting to select a distance measure, the researcher should remember the following caveats. Different distance measures or changing the scales of the variables may lead to different cluster solutions. Thus, it is advisable to use several measures and compare the results with theoretical or known patterns. Also when the variables are intercorrelated (either positively or negatively), the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for intercorrelations and weights all variables equally. Of course, if the researcher wishes to weight the variables unequally, other procedures are available (Overall, 1964).

Association measures. Association measures of similarity are used to compare objects whose characteristics are measured only in nonmetric terms (nominal or ordinal measurement). As an example, respondents could answer yes or no on a number of statements. An association measure could assess the degree of agreement or matching between each pair of respondents. The simplest form of association measure would

be the percentage of times there was agreement (both respondents answered yes or both said no to a question) across the set of questions. Extensions of this simple matching coefficient have been developed to accommodate multicategory nominal variables and even ordinal measures. Many computer programs, however, have limited support for association measures, and the researcher is often forced to first calculate the similarity measures and then input the similarity matrix into the cluster program. Reviews of the various types of association measures can be found in several sources (Everitt, 1980; Sneath & Sokal, 1973).

Data Standardization

With the similarity measure selected, the researcher must address only one more question: Should the data be standardized before similarities are calculated? In answering this question, the researcher must address several issues. First, most distance measures are sensitive to differing scales or magnitude among the variables. We saw this effect earlier when we changed from minutes to seconds on one of our variables. In general, variables with larger dispersion (i.e., larger standard deviations) have a greater effect on the final similarity value. Consider another example to illustrate this point. Assume that we want to cluster individuals on three variables: an attitude toward a product, age, and income. Now assume that we measure attitude on a 7-point scale of *liking* (7) to *disliking* (1), whereas age is measured in years, and income in dollars. If we plotted this on a three-dimensional graph, the distance between points (and their similarity) would be almost totally based on the income differences. The possible differences in attitude range from one to seven, whereas income may have a range perhaps a thousand times greater. Thus, graphically we would not be able to see any difference on the dimension associated with attitude. For this reason, the researcher must be aware of the implicit weighting of variables based on their relative dispersion that occurs with distance measures.

Standardization by variables. The most common form of standardization is the conversion of each variable to standard scores (also known as "z scores") by subtracting the mean and dividing by the standard deviation for each variable. This is an option in all computer programs, and many times it is even directly included in the cluster analysis procedure. This is the general form of a *normalized distance function*, which uses a Euclidean distance measure amenable to a normalizing transformation of the raw data. This process converts each raw data score into a standardized value with a zero mean and a unit standard deviation.

This transformation, in turn, eliminates the bias introduced by the differences in the scales of the several attributes or variables used in the analysis.

The benefits of standardization can be seen in Table 5.4D, where two variables (purchase probability and viewing time) have been standardized before computing the three distance measures. First, it is much easier to compare between the variables because they are on the same scale (a mean of 0 and standard deviation of 1). Positive values are above the mean and negative values below, with the magnitude representing the number of standard deviations the original value was from the mean. Second, there is no difference in the standardized values when only the scale changes. For example, when viewing time in minutes and then seconds is standardized, the values are the same. Thus, using standardized variables truly eliminates the effects due to scale differences not only across variables but also for the same variable. The researcher should not always apply standardization without consideration for its consequences, however. There is no reason to absolutely accept the cluster solution using standardized variables versus unstandardized variables. If there is some natural relationships reflected in the scaling of the variables, then standardization may not be appropriate. One example of a natural relationship would be a study that uses both overall and specific measures where the researcher wants the overall measure to be attributed a greater weight. For example, the study may use 5-point measures of job satisfaction for specific attributes and a 10-point measure for life satisfaction. In this study, due to the natural relationship, the researcher would not want to standardize the variables. The decision to standardize has both empirical and conceptual implications and should always be made with careful consideration.

Standardization by observation. What about standardizing respondents or cases? Why would one ever do this? For example, suppose that we had collected a number of ratings on a 10-point scale from respondents on the importance of several attributes in their purchase decision for a product. We could apply cluster analysis and obtain clusters, but one distinct possibility is that we would get clusters of people who said everything was important, some who said everything had little importance, and perhaps some clusters in between. This is in the clusters are *response-style effects*, which are the systematic patterns of responding to a set of questions, such as yea-sayers (those who answer very favorably to all questions) or nay-sayers (those who answer unfavorably to all questions).

If we want to identify groups according to their response style, then standardization is not appropriate. But in most instances, what is desired is the relative importance of one variable to another. In other words, is attribute 1 more or less important than the other attributes? Can clusters of respondents be found with similar patterns of importance? In this instance, standardizing by respondent would standardize each question not to the sample's average but instead to that respondent's average score. This *within-case* or *row-centering standardization* can be effective in removing response effects, and it is especially suited to many forms of attitudinal data. This is similar to a correlational measure in highlighting the pattern across variables, but the proximity of cases still determines the similarity value.

Stage 3: Assumptions

Cluster analysis, like multidimensional scaling, is not a statistical inference technique in which parameters from a sample are assessed as possibly being representative of a population. Instead, cluster analysis is an objective methodology for quantifying the structural characteristics of a set of observations. As such, it has strong mathematical properties, not statistical foundations. The requirements of normality, linearity, and homoscedasticity that are so important in other techniques have little bearing on cluster analysis. The researcher must focus, however, on two other critical issues: (a) representativeness of the sample and (b) multicollinearity.

Representativeness of the Sample

In very few instances does the researcher have a census of the population to use in the cluster analysis. Instead, a sample of cases is obtained and the clusters derived in the hope that they represent the structure of the population. The researcher must therefore be confident that the obtained sample is truly representative of the population. As mentioned earlier, outliers may really be only an undersampling of divergent groups that when discarded, introduce bias in the estimation of structure. The researcher must realize that cluster analysis is only as good as the representativeness of the sample. Therefore, all efforts should be taken to ensure that the sample is representative and that the results are generalizable to the population of interest.

Multicollinearity

Multicollinearity is an issue in other multivariate techniques (e.g., multiple regression) because it makes it difficult to discern the true effect

of multicollinear variables. In cluster analysis, however, the effect is different because those variables that are multicollinear are implicitly weighted more heavily. Let us start with an example that illustrates its effect. Suppose that respondents are being clustered on 10 variables, all attitudinal statements concerning a service. When multicollinearity is examined, we see that there are really two sets of variables: the first made up of eight statements and the second consisting of the remaining two statements. If our intent is to really cluster the respondents on the dimensions of the product (in this case represented by the two groups of variables), then use of the original 10 variables would be misleading. Because each variable is weighted equally in cluster analysis, the first dimension has four times as many chances (eight items to two items) to affect the similarity measure as does the second dimension.

Multicollinearity acts as a weighting process not apparent to the observer but affecting the analysis nonetheless. For this reason, the researcher is encouraged to examine the variables used in cluster analysis for substantial multicollinearity and, if found, to either reduce the variables to equal numbers in each set or use one of the distance measures, such as Mahalanobis distance, that compensates for this correlation. There is debate over the use of factor scores in cluster analysis, as some research has shown that the variables that truly discriminate among the underlying groups are not well represented in most factor solutions. Thus, when factor scores are used, it is possible that a poor representation of the true structure of the data is obtained (Schaninger & Bass, 1986). The researcher must deal with both multicollinearity and discriminability of the variables to arrive at the best representation of structure.

Stage 4: Derivation of Clusters and Assessment of Overall Fit

With the variables selected and the similarity matrix calculated, the partitioning process begins. The researcher must first select the clustering algorithm used for forming clusters and then make the decision on the number of clusters to be formed. Both decisions have substantial implications not only on the results that are obtained but also on the interpretation that can be derived from the results. Each issue is discussed in the following sections. Figure 5.6 illustrates Stages 4–6 of the cluster analysis decision tree.

Figure 5.6

Stage 4

Hiera
Linkage
Single
Compl
Averag
Ward's
Centroi

Stage 5

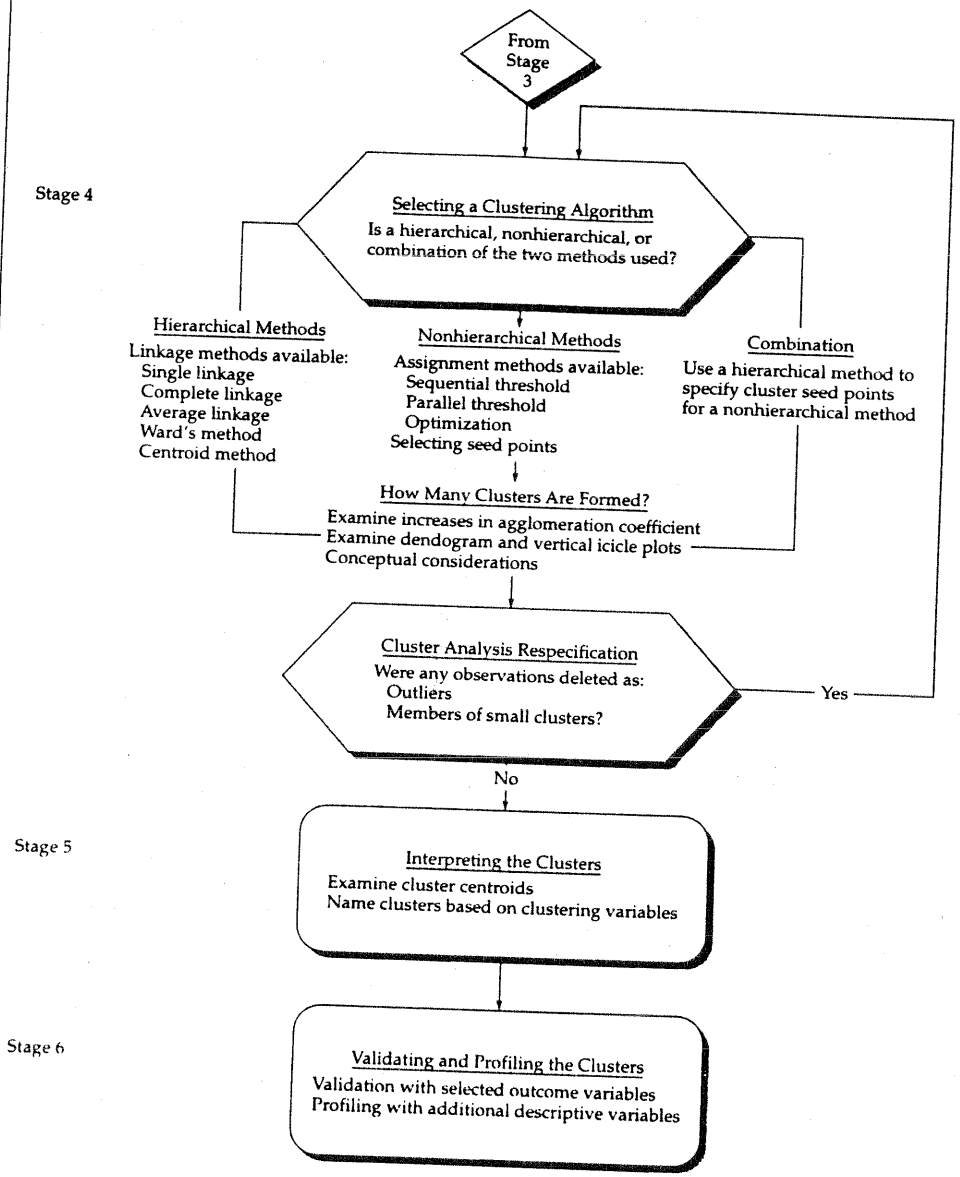
Stage 6

Stages 1–6 of th

Clustering Algorit

The first major
procedure shou
ters?" That is,

Figure 5.6

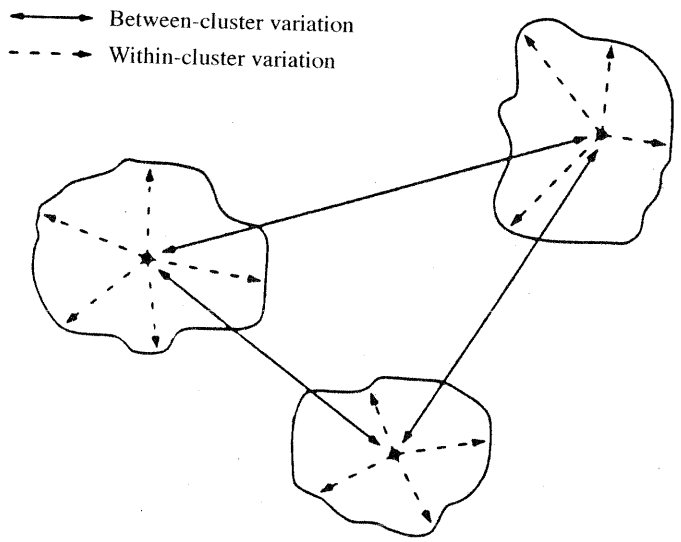


Stages 1–6 of the cluster analysis decision diagram.

Clustering Algorithms

The first major question to answer in the partitioning phase is, "What procedure should be used to place similar objects into groups or clusters?" That is, what clustering algorithm or set of rules is the most

Figure 5.7



Cluster diagram showing between- and within-cluster variation.

appropriate? This is not a simple question because hundreds of computer programs using different algorithms are available, and more are always being developed. The essential criterion of all the algorithms, however, is that they attempt to maximize the differences between clusters relative to the variation within the clusters, as shown in Figure 5.7. The ratio of the between-cluster variation to the average within-cluster variation is then comparable to (but not identical to) the F ratio in analysis of variance. Most commonly used clustering algorithms can be classified into two general categories: (a) hierarchical and (b) nonhierarchical.

Hierarchical Cluster Procedures

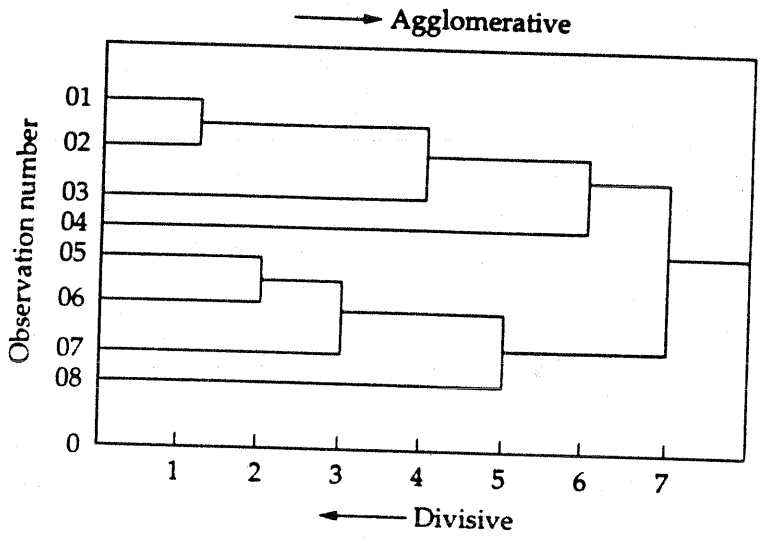
Hierarchical procedures involve the construction of a hierarchy of a treelike structure. There are basically two types of hierarchical clustering procedures: (a) agglomerative and (b) divisive. In the agglomerative methods, each object or observation starts out as its own cluster. In subsequent steps, the two closest clusters (or individuals) are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. In some cases, a third individual joins the first two in a cluster. In others, two groups of individuals formed at an earlier stage

may join together in a new cluster. Eventually, all individuals are grouped into one large cluster; for this reason, agglomerative procedures are sometimes referred to as "build-up methods."

An important characteristic of hierarchical procedures is that the results from an earlier stage are always nested within the results in a later stage, creating its similarity to a tree. For example, a six-cluster solution is obtained by joining two of the clusters found at the seven-cluster stage. Because clusters are formed only by joining existing clusters, any member of a cluster can trace its membership in an unbroken path to its beginning as a single observation. This process is shown in Figure 5.8; the representation is referred to as a *dendrogram* or *tree graph.*

When the clustering process proceeds in the opposite direction to agglomerative methods, it is referred to as a *divisive method.* In divisive methods, we begin with one large cluster containing all of the observations (objects). In succeeding steps, the observations that are most dissimilar are split off and made into smaller clusters. This process continues until each observation is a cluster in itself. In Figure 5.8, agglomerative methods would move from left to right, and divisive methods would move from right to left. Because most commonly used computer

Figure 5.8



Dendogram illustrating hierarchical clustering.

packages use agglomerative methods, and divisive methods act almost as agglomerative methods in reverse, we focus on the agglomerative methods in our subsequent discussions.

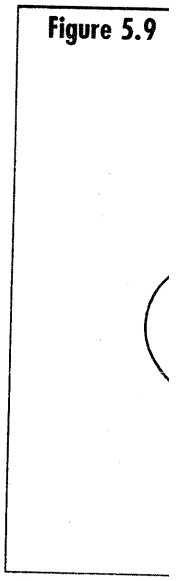
Five popular agglomerative algorithms used to develop clusters are (a) single linkage, (b) complete linkage, (c) average linkage, (d) Ward's method, and (e) centroid method. These rules differ in how the distance between clusters is computed.

Single linkage. *Single linkage* is based on minimum distance. It finds the two objects separated by the shortest distance and places them in the first cluster. Then the next-shortest distance is found, and either a third object joins the first two to form a cluster or a new two-member cluster is formed. The process continues until all objects are in one cluster. This procedure has also been called the "nearest-neighbor approach."

The distance between any two clusters is the shortest distance from any point in one cluster to any point in the other. Two clusters are merged at any stage by the single shortest or strongest link between them. This was the rule applied in the example at the beginning of the chapter. Problems occur, however, when clusters are poorly delineated. In such cases, single linkage procedures can form long, snakelike chains, and eventually all individuals are placed in one chain. Individuals at opposite ends of a chain may be very dissimilar.

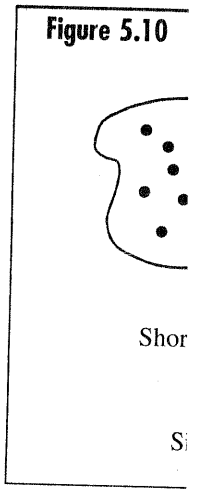
An example of this arrangement is shown in Figure 5.9. Three clusters (A, B, and C) are to be joined. The single-linkage algorithm, focusing on only the closest points in each cluster, would link Clusters A and B because of their short distance at the extreme ends of the clusters. Joining Clusters A and B creates a cluster that encircles Cluster C. Yet in striving for within-cluster homogeneity, it would be much better to join Cluster C with either A or B. This is the principal disadvantage of the single-linkage algorithm.

Complete linkage. *Complete linkage* is similar to single linkage except that the cluster criterion is based on maximum distance. For this reason, it is sometimes referred to as the "furthest-neighbor approach" or a "diameter method." The maximum distance between individuals in each cluster represents the smallest (minimum-diameter) sphere that can enclose all objects in both clusters. This method is complete because all of the objects in a cluster are linked to each other at some maximum distance or by minimum similarity. We can say that within-group similarity equals group diameter. This technique eliminates the snaking problem identified with single linkage.



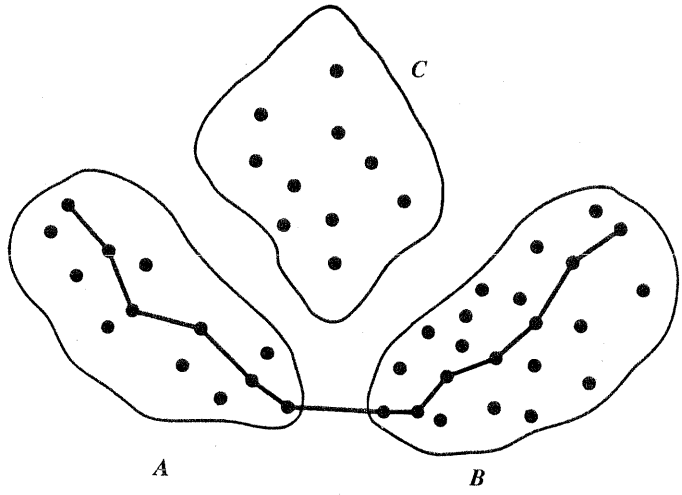
A single link

Figure
(complete li
measures ref
distance refl
complete lin



Comparison o

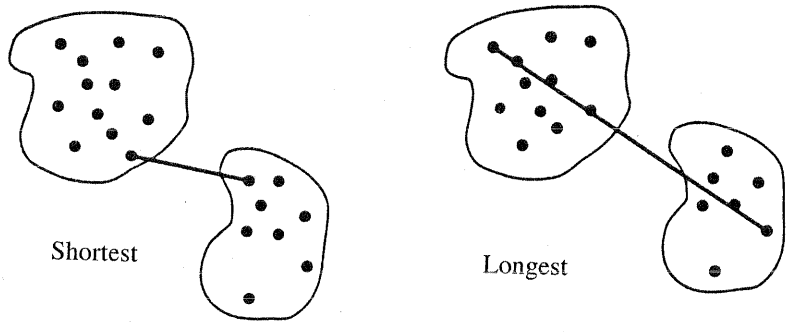
Figure 5.9



A single linkage joining dissimiliar Clusters A and B.

Figure 5.10 shows how the shortest (single linkage) and longest (complete linkage) distances represent similarity between groups. Both measures reflect only one aspect of the data. The use of the shortest distance reflects only a single pair of objects (the closest), while the complete linkage again reflects only a single pair, but this time it is the

Figure 5.10



Comparison of distance measures for single linkage and complete linkage.

two most extreme. It is thus useful to visualize the measures as reflecting the similarity of most similar pair or least similar pair of objects.

Average linkage. *Average linkage* starts out the same as a single or a complete linkage, but the cluster criterion is the average distance from all individuals in one cluster to all individuals in another. Such techniques do not depend on extreme values, as do single linkage or complete linkage, and partitioning is based on all members of the clusters rather than on a single pair of extreme members. Average-linkage approaches tend to combine clusters with small within-cluster variation. They also tend to be biased toward the production of clusters with approximately the same variance.

Ward's method. In *Ward's method,* the distance between two clusters is the sum of squares between the two clusters summed over all variables. At each stage in the clustering procedure, the within-cluster sum of squares is minimized over all partitions (the complete set of disjoint or separate clusters) obtainable by combining two clusters from the previous stage. This procedure tends to combine clusters with a small number of observations. It is also biased toward the production of clusters with approximately the same number of observations.

Centroid method. In the *centroid method,* the distance between two clusters is the distance (typically squared Euclidean or simple Euclidean) between their centroids. *Cluster centroids* are the mean values of the observations on the variables in the cluster variate. In this method, every time individuals are grouped, a new centroid is computed. Cluster centroids migrate as cluster mergers take place. In other words, there is a change in a cluster centroid every time a new individual or group of individuals is added to an existing cluster. This method is the most popular with biologists, but it may produce messy and often confusing results. The confusion occurs because of reversals; that is, instances when the distance between the centroids of one pair may be less than the distance between the centroids of another pair merged at an earlier combination. The advantage of this method is that it is less affected by outliers than are other hierarchical methods.

Nonhierarchical Clustering Procedures

In contrast to hierarchical methods, *nonhierarchical procedures* do not involve the treelike construction process. Instead, they assign objects into clusters once the number of clusters to be formed is specified. Thus, the six-cluster solution is not just a combination of two clusters from

the seven-cluster solution but is based only on finding the best six-cluster solution. In a simple example, the process works this way. The first step is to select a cluster seed as the initial cluster center, and all objects (individuals) within a prespecified threshold distance are then included in the resulting cluster. Then, another cluster seed is chosen and the assignment continues until all objects are assigned. Then, objects may be reassigned if they are closer to another cluster than the one originally assigned. There are several approaches for selecting cluster seeds and assigning objects, which we discuss in the next section. Nonhierarchical clustering procedures are frequently referred to as "K-means clustering," and they typically use one of the following three approaches for assigning individual observations to one of the clusters (Green, 1978, p. 428).

Sequential threshold. The *sequential threshold method* starts by selecting one cluster seed and including all objects within a prespecified distance. When all objects within the distance are included, a second cluster seed is selected, and all objects within the prespecified distance are included in that cluster. Then a third seed is selected, and the process continues as before. When an object is clustered with a seed, it is no longer considered for subsequent seeds.

Parallel threshold. In contrast, the *parallel threshold method* selects several cluster seeds simultaneously in the beginning and assigns objects within the threshold distance to the nearest seed. As the process evolves, threshold distances can be adjusted to include fewer or more objects in the clusters. Also, in some methods, objects remain unclustered if they are outside the prespecified threshold distance from any cluster seed.

Optimization. *Optimizing procedure* is similar to the other two except that it allows for the reassignment of objects. If in the course of assigning objects, an object becomes closer to another cluster that is not the cluster to which it was originally assigned, then an optimizing procedure switches the object to the more similar (closer) cluster.

Seed point selection. Nonhierarchical procedures are available in a number of computer programs, including all of the major statistical packages. The sequential threshold procedure (e.g., FASTCLUS program in SAS) is an example of a nonhierarchical clustering program designed for large data sets. After the researcher specifies the maximum number of clusters allowed, the procedure begins by selecting cluster seeds, which are used as initial guesses of the means of the clusters. The first seed is the first observation in the data set with no missing values. The second seed is the next complete observation (no missing data)

that is separated from the first seed by a specified minimum distance. The default option is a zero minimum distance. After all seeds have been selected, the program assigns each observation to the cluster with the nearest seed. The researcher can specify that the cluster seeds be revised (updated) by calculating seed cluster means each time an observation is assigned. In contrast, the parallel threshold methods (e.g., QUICK CLUSTER in SPSS) establish the seed points as user-supplied points or select them randomly from all observations.

The major problem faced by all nonhierarchical clustering procedures is how to select the cluster seeds. For example, with a sequential threshold option, the initial and probably the final cluster results depend on the order of the observations in the data set, and shuffling the order of the data is likely to affect the results. Specifying the initial cluster seeds as in the sequential threshold procedure can reduce this problem. Even selecting the cluster seeds randomly, however, produces different results for each set of random seed points. Thus, the researcher must be aware of the effect of the cluster seed selection process on the final results.

Hierarchical Versus Nonhierarchical Methods

A definitive answer to the question of whether hierarchical or nonhierarchical methods should be used cannot be given for two reasons. First, the research problem at hand typically may suggest one method or the other. Second, what we learn with continued application of one method to a particular context may suggest one method over the other as more suitable for that context.

Pros and cons of hierarchical methods. In the past, hierarchical clustering techniques were more popular, with Ward's method and average linkage probably the best available (Milligan, 1980). Hierarchical procedures do have the advantage of being fast and, therefore, taking less computer time. Yet with the computing power of today, even personal computers can handle large data sets easily. Hierarchical procedures can be misleading, however, because undesirable early combinations may persist throughout the analysis and lead to artificial results. Of concern is the substantial effect of outliers on hierarchical methods, particularly with the complete linkage method. To reduce this possibility, the researcher may wish to cluster analyze the data several times, each time deleting problem observations or outliers. The deletion of cases, even those not found to be outliers, can many times distort the

solution. Thus, the researcher must use extreme care in the deletion of observations for any reason.

Also, although computations of the clustering process are relatively fast, hierarchical methods are not amenable to analyzing very large samples. As sample size increases, the data storage requirements increase dramatically. For example, a sample of 400 cases requires storage of approximately 80,000 similarities, and this number increases to almost 125,000 for a sample of 500. Even given today's technological advances, problems of this size exceed the capacity of most personal computers, thus limiting the application of hierarchical methods in many instances. The researcher may take a random sample of the original observations to reduce its size but must now question the representativeness of the sample taken from the original sample.

Emergence of nonhierarchical methods. Nonhierarchical methods have gained acceptability and are applied increasingly. Their use, however, depends on the ability of the researcher to select the seed points according to some practical, objective, or theoretical basis. In these instances, nonhierarchical methods have several advantages over hierarchical techniques. The results are less susceptible to the outliers in the data, the distance measure used, and the inclusion of irrelevant or inappropriate variables. These benefits are realized, however, only with the use of nonrandom (i.e., specified) seed points; thus, the use of nonhierarchical techniques with random seed points is markedly inferior to the hierarchical techniques. Even a nonrandom starting solution does not guarantee an optimal clustering of observations. In fact, in many instances, the researcher gets a different final solution for each set of specified seed points. How is the researcher to select the "correct" answer? Only by analysis and validation can the researcher then select what is considered the "best" representation of structure, realizing that there are many alternatives that may be as acceptable.

A combination of both methods. Another approach is to use both methods (hierarchical and nonhierarchical) to gain the benefits of each (Milligan, 1980). First, a hierarchical technique can establish the number of clusters, profile the cluster centers, and identify any obvious outliers. After outliers are eliminated, the remaining observations are then clustered by a nonhierarchical method, with the cluster centers from the hierarchical results as the initial seed points. In this way, the advantages of the hierarchical methods are complemented by the ability of the nonhierarchical methods to "fine-tune" the results by allowing the switching of cluster membership.

Number of Clusters Formed

Perhaps the most perplexing issue for a researcher using cluster analysis is determining the final number of clusters to be formed (also known as the "stopping rule"). Unfortunately, no standard, objective selection procedure exists. Because there is no internal statistical criterion used for inference, such as the statistical significance tests of other multivariate methods, researchers have developed many criteria and guidelines for approaching the problem. The principal drawback is that these are ad hoc procedures that must be computed by the researcher, and many times they involve fairly complex procedures (Aldenderfer & Blashfield, 1984; Milligan & Cooper, 1985). One class of stopping rules that is relatively simple examines some measure of similarity or distance between clusters at each successive step, with the cluster solution defined when the similarity measure exceeds a specified value or when the successive values between steps makes a sudden jump. A simple example of this was used in the example at the beginning of the chapter, which looked for large increases in the average within-cluster distance. When a large increase occurs, the researcher selects the prior cluster solution on the logic that its combination caused a substantial decrease in similarity. This stopping rule has been shown to provide fairly accurate decisions in empirical studies (Milligan & Cooper, 1985). A second general class of stopping rules attempts to apply some form of statistical rule or to adapt a statistical test, such as the point-biserial/tau correlations or the likelihood ratio. Although some of these have been shown to have notable success, such as the cubic clustering criterion contained in SAS, many seem overly complex for the improvement that they provide over simpler measures. There are other specific procedures that have been proposed, but none have been found to be substantially better in all situations.

The researcher should also complement the strictly empirical judgment with any conceptualization of theoretical relationships that may suggest a natural number of clusters. Also one might start this process by specifying some criteria on the basis of practical considerations, such as saying "My findings will be more manageable and easier to communicate if I have between three and six clusters" and then solving for this number of clusters and selecting the best alternative after evaluating all of them. In the final analysis, however, it is probably best to compute a number of cluster solutions (e.g., two, three, and four) and then decide among the alternative solutions by using a priori criteria, practical judgment, common sense, or theoretical foundations. The cluster so-

lutions are improved by restricting the solution according to conceptual aspects of the problem.

Respecification of the Cluster Analysis

When an acceptable cluster analysis solution is identified, the researcher should examine the fundamental structure represented in the defined clusters. Of note are widely disparate cluster sizes or clusters of only one or two observations. Researchers must examine widely varying cluster sizes from a conceptual perspective, comparing the actual results with the expectations formed in the research objectives. More troublesome are single-member clusters, which may be outliers not detected in earlier analyses. If a single-member cluster (or one of very small size compared with other clusters) appears, the researcher must decide if the cluster represents a valid structural component in the sample or if it should be deleted as unrepresentative. If any observations are deleted, especially when hierarchical solutions are used, the researcher should rerun the cluster analysis and start the process of defining clusters anew.

Stage 5: Interpretation of the Clusters

The interpretation stage involves examining each cluster in terms of the cluster variate to name or assign a label accurately describing the nature of the clusters. To clarify this process, let us refer to the example of diet versus regular soft drinks. Assume that an attitude scale was developed that consisted of statements regarding consumption of soft drinks. Examples of statements include "Diet soft drinks taste harsher," "Regular soft drinks have a fuller taste," "Diet drinks are healthier." Assume further that demographic and soft drink consumption data were also collected.

When starting the interpretation process, one measure frequently used is the cluster's centroid. If the clustering procedure were performed on the raw data, this would be a logical description. If the data were standardized or if the cluster analysis were performed using factor analysis (component factors), the researcher would have to go back to the raw scores for the original variables and compute average profiles using these data.

Continuing with the soft drink example, in this stage we examine the average score profiles on the attitude statements for each group and assign a descriptive label to each cluster. Many times discriminant analysis is applied to generate score profiles, but we must remember that statistically significant differences would not indicate an "optimal"

solution because statistical differences are expected given the objective of cluster analysis. Examination of the profiles allows for a rich description of each cluster. For example, two of the clusters may have favorable attitudes about diet soft drinks, and the third cluster may have negative attitudes. Moreover, of the two favorable clusters, one may exhibit favorable attitudes toward only diet soft drinks, whereas the other may display favorable attitudes toward both diet and regular soft drinks. From this analytical procedure, one would evaluate each cluster's attitudes and develop substantive interpretations to facilitate labeling each. For example, one cluster might be labeled "Health and calorie conscious," whereas another might be "Get a sugar rush."

The profiling and interpretation of the clusters, however, achieve more than just description. First, they provide a means for assessing the correspondence of the derived clusters to those proposed by prior theory or practical experience. If used in a confirmatory mode, the cluster analysis profiles provide a direct means of assessing the correspondence. Second, the cluster profiles provide a route for making assessments of practical significance. The researcher may require that substantial differences exist on a set of clustering variables and the cluster solution is expanded until such differences arise. In either instance, the researcher has a plan for comparing the derived clusters to a preconceived typology.

Stage 6: Validation and Profile of the Clusters

Given the somewhat subjective nature of cluster analysis with regard to selecting an "optimal" cluster solution, the researcher should take great care in validating and ensuring the practical significance of the final cluster solution. Although no single method exists to ensure validity and practical significance, several approaches have been proposed to provide some basis for the researcher's assessment.

Validation of the Cluster Solution

Validation includes attempts by the researcher to ensure that the cluster solution is representative of the general population and, thus, is generalizable to other objects and stable over time. The most direct approach in this regard is to cluster analyze separate samples, comparing the cluster solutions and assessing the correspondence of the results. This approach, however, is often impractical because of time or cost constraints or the unavailability of research participants for multiple cluster analyses. In these instances, a common approach is to split the sample into

two groups. Each is cluster analyzed separately, and the results are then compared. Other approaches include (a) a modified form of split sampling whereby cluster centers obtained from one cluster solution are used to define clusters from the other observations and the results are compared (McIntyre & Blashfield, 1980) and (b) a direct form of cross-validation (Punj & Stewart, 1983).

The researcher may also attempt to establish some form of criterion or predictive validity. To do so, the researcher selects variables not used to form the clusters but known to vary across the clusters. In our example, we may know from past research that attitudes toward diet soft drinks vary by age. Thus, we can statistically test for the differences in age between those clusters favorable to diet soft drinks and those that are not. The variables used to assess predictive validity should have strong theoretical or practical support because they become the benchmark for selecting among the cluster solutions.

Profile of the Cluster Solution

Profiling involves describing the characteristics of each cluster to explain how they may differ on relevant dimensions. This typically involves the use of discriminant analysis or some other appropriate statistic that tests for differences among means. The procedure begins after the clusters are identified. The researcher uses data not previously included in the cluster procedure to profile the characteristics of each cluster. These data typically are demographic characteristics, psychological profiles, consumption patterns, and so forth. Although there may not be a theoretical rationale for their difference across the clusters, such as is required for predictive validity assessment (see above), the data should at least have practical importance. Using discriminant analysis or analysis of variance, the researcher can then compare average score profiles for the clusters. The categorical dependent variable is the previously identified clusters, and the independent variables are the demographics, psychographics, and so on. From this analysis, assuming statistical significance, the researcher could conclude, for example, that the "Health and calorie conscious" cluster from our previous example consists of better educated, higher income professionals who are moderate consumers of soft drinks. In short, the profile analysis focuses on describing not what directly determines the clusters, but the characteristics of the clusters after they have been identified. Moreover, the emphasis is on the characteristics that differ significantly across the clusters and those that could predict membership in a particular attitude cluster.

An Example of Stages: Who Are the Homeless?

We provide another example of the use of cluster analysis. In the 1980s, citizens of the United States became increasingly aware of and sensitive to the plight of the homeless. It was not long before social scientists began to gather data in an attempt to identify the characteristics of the homeless population, embodied by the question "Who are the homeless?" This is the type of question that lends itself to cluster analysis because the analysis can identify naturally occurring groups, or types of homeless people. Within the context of the following example, we identify how the researchers addressed some of the issues that we present in our organizational schema, or stage model of cluster analysis.

Mowbray, Bybee, and Cohen (1993) identified 108 homeless people through various community agencies and shelters. Information about the research participants was obtained from agency referral forms, client interviews, ratings of outreach workers, and archival records. Because the researchers were interested in "dimensions thought to be important for community functioning," they were guided in the selection of measures; previous research had suggested good candidates: community living problems, depression, substance abuse, psychoticism, and aggression.

- *Stage 1: Objectives.* The authors used cluster analysis to establish a taxonomy of the homeless population. In addition, despite the exploratory nature of their study, variables were not selected arbitrarily; their selection was based on previous research that established a relationship between these variables and community functioning. Moreover, because the sample size was relatively small, a decision was made to limit the number of variables included in the analysis.
- *Stage 2: Research Design.* The researchers examined the data for outliers. Because, in their original format, the variables were scaled differently, a situation that can lead to undo weighting of scales with large ranges, the authors standardized the scales before submitting the data for analysis. In addition, they chose to use Euclidean distances to calculate interindividual similarities.
- *Stage 3: Assumptions.* The researchers did their best to obtain a representative sample. Although all participants came from one Midwestern state, about half the sample came from an urban

setting and half from a college town. About 50% of the clients were recruited by referrals from community mental health centers, 30% were recruited from psychiatric inpatient facilities, and 17% of the sample were living in shelters and had no records at the community health centers. The total sample had a large number of people who had current or past involvement with the mental health system. For this reason, the title of the researchers' article appropriately includes the wording "homeless mentally ill."

Another issue in cluster analysis is *multicollinearity*, the intercorrelation among variables used in clustering. Variables that are too highly correlated with one another distort the cluster solution by being overweighted. Of the 10 correlations in the matrix, there was a significant correlation (+.47) between community living problems and psychoticism. Because the correlation was moderate and the variables were conceptually distinct, the authors retained both variables for inclusion in the analysis.

The results of the analysis suggested the viability of three or four clusters. In examining the characteristics of the clusters, defined by only those variables used in the cluster analysis, the authors chose names for each cluster that best captured the meaning of the variables. The first cluster was "Hostile–psychotic" and included 35% of the sample. These clients scored the highest on measures of aggression and psychoticism. The second-largest cluster was "Best functioning" and included 28% of the sample. Clients in this group had low scores on all measures of dysfunction. In the identification of the third and fourth clusters, a decision had to be made. When viewed together as one cluster, the most apt name for this group was "Depressed." However, the results of the analysis suggested that a fourth cluster may have meaning. The Depressed group included individuals who abused substances and those who did not. In the authors' words, "Inspection of the clusters defined by each solution revealed significant and theoretically meaningful differences between the fourth cluster and the other three" (Mowbray et al., 1993, p. 86). Consequently, the Depressed group that was retained included 19% of the sample and was defined by high scores on depression and low scores on substance abuse. The fourth cluster, "Substance abusing," made up 19% of the total sample and was de-

fined by high scores on substance abuse and the second-highest scores on depression.

- *Stage 4: Derivation of Clusters and Assessment of Overall Fit.* In deriving clusters, the researchers used a combination of hierarchical (Ward's method) and nonhierarchical (K means) clustering, with the centroids of the first method used as seed points in the second cluster analysis. Because the sample size was relatively small, the researchers selected only five variables for inclusion to guard against over fitting.
- *Stage 5: Interpretation of the Clusters.* Interpreting clusters involves examining the differences between the clusters with respect to the variables used in the cluster analysis. At first glance, one might think that because the analysis separated clusters on the basis of these variables, there would be a significant difference between clusters on all variables. This is rarely the case. The authors conducted univariate *F* tests on each variable and used the pattern of significant differences to name the clusters. For example, the Substance-abusing cluster reported significantly greater substance abuse than the other clusters, but there were no differences among the other clusters on this variable. Likewise, the Hostile–psychotic cluster showed high scores on hostility and psychoticism, which were significantly different than the other clusters, but no differences on these variables were observed among the remaining three clusters.

  The authors compared the groups on other variables not included in the cluster analysis. For example, the people in the Hostile–psychotic cluster scored higher on deviancy and showed the worst overall functioning; the Depressed group had the highest potential for suicide; women tended to be overrepresented in the Depressed group; and the Substance-abusing cluster tended to have younger people, while the Hostile–psychotic group tended to have more people over the age of 40.
- *Stage 6: Validation and Profile of the Clusters.* The authors stressed the profiling of clusters by comparing the groups on variables external to the cluster analysis. Moreover, they chose variables that would be expected to differ across the clusters. One outcome of the cluster analysis and profiling stage was that it may be possible to offer a narrative description of the characteristics of the members of the cluster.

Summary of the Decision Process

Cluster analysis provides researchers with an empirical and objective method for performing one of the most inherent tasks for humans: classification. Whether for purposes of simplification, exploration, or confirmation, cluster analysis is a potent analytical tool that has a wide range of applications. With this technique, however, comes a responsibility on the part of the researcher to apply the underlying principles appropriately. As mentioned in the beginning of this chapter, cluster analysis has many caveats that cause even the experienced researcher to apply it with caution. Yet when used appropriately, it has the potential to reveal structures within the data that could not be discovered by any other means. Thus, this powerful technique addresses a fundamental need of researchers in all fields with the knowledge that is can be as easily abused as used wisely.

An Example of An Application: The HATCO Database

To illustrate the application of cluster analysis techniques, I turn to the HATCO database.[1] The seven perceptions of HATCO provide a basis for illustrating one of the most common uses of cluster analysis: the formation of customer segments. In our example, we follow the sequence of stages (not enumerated) of the model-building process presented earlier in the chapter. The seven HATCO perceptions, or attributes, rated by each respondent included delivery speed, price, price flexibility, manufacturer's image, overall service, sales force image, and product quality.

We begin by cluster analyzing the ratings of HATCO customers as to the performance of HATCO on the seven attributes (X_1 to X_7). Our objective is to segment objects (customers) into groups with similar perceptions of HATCO. Once identified, HATCO can then formulate strat-

[1]HATCO stands for the Hair, Anderson, and Tatham Company, a large (although nonexistent) industrial supplier. The data set was obtained from a survey of 100 HATCO customers, assessed on 14 separate variables, collected through an established marketing research firm. Three types of information were collected. The first type is the perception of HATCO on seven attributes, identified in past studies as the most influential in the choice of suppliers. The respondents, purchasing managers of firms buying from HATCO, rated HATCO on each attribute. The second type of information relates to actual purchase outcomes, either the evaluations of each respondent's satisfaction with HATCO or the percentage of that respondent's product purchases from HATCO. The third type of information contains general characteristics of the purchasing companies, recorded as nonmetric data (e.g., firm size, or industry type).

egies with different appeals for the separate groups—the requisite basis for market segmentation. A primary concern is that the seven attributes used to form the clusters be adequate in scope and detail. From the examples in other chapters with the various multivariate techniques, we have found that these variables have sufficient predictive power to justify their use as the basis for segmentation. The sample of 100 observations was examined for outliers and was found to have no strong candidates for deletion (one should examine the cluster solutions in later stages and assess if outliers have emerged during the clustering process). Given that the set of seven variables is metric, squared Euclidean distance was chosen as the similarity measure. Standardization of variables was not undertaken because all variables were on the same scale, and within-case standardization was not appropriate because the magnitude of the perceptions was an important element of the segmentation objectives. The sample was considered a representative sample of HATCO customers. An analysis of multicollinearity identified only minimal levels that should not influence the cluster analysis in any substantial manner.

Ward's method was chosen to minimize the within-cluster differences and to avoid problems with "chaining" of the observations found in the single-linkage method. Because the data involve profiles of HATCO customers and our interest is in identifying types or profiles of these customers that may form the bases for differing strategies, a manageable number of clusters was deemed to be in the range of two to five. The clustering (agglomeration) coefficient showed rather large increases in going from four to three clusters, three to two clusters, and two to one cluster. The largest percentage of increase in the clustering coefficient occurred in going from two to one cluster, the next noticeable change in the percentage of increase occurred in combining four into three clusters. Thus, both the two- and four-cluster solutions were examined.

Table 5.5 contains the clustering variable profiles for both cluster solutions. An examination of the two-cluster profiles reveals two clusters that are almost mirror images of each other. Cluster 1 has high values on X_1, X_3, and X_5, whereas Cluster 2 has higher values for X_2, X_4, X_6, and X_7. Also shown in Table 5.5 are the profiles for the four-cluster solution, which reveal a number of patterns of high versus low values. Another aspect that varies from the two-cluster solution is that all of the clustering variables vary in a statistically significant manner across the four groups, versus only five of the variables in the two-cluster solution. Because all indicators (stopping rule, absence of outliers, and distinctive

profiles) supported either the two- or four-cluster solutions, both solutions were carried forward into the nonhierarchical analysis to obtain final cluster solutions. The results (centroid values and cluster size) of the nonhierarchical are shown in Table 5.6 for both cluster solutions. As seen in the hierarchical methods, the two-cluster solution results in groups of almost equal size (52 vs. 48 observations) and the profiles correspond well with the cluster profiles from the hierarchical procedure. Again, only X_5 shows no significant differences between the clusters. For the four-group solution, the cluster sizes are similar to those from the hierarchical procedure, varying in size at the most by only four observations. Also the cluster profiles match well. The correspondence and stability of the two solutions between the nonhierarchical and hierarchical methods confirms the results subject to theoretical and practical acceptance.

Information essential to the interpretation and profiling stages is provided in Table 5.6. For each cluster, the centroid on each of the seven clustering variables is provided, along with the univariate *F* ratios and levels of significance comparing the differences between the cluster means. Factor analysis revealed that two factors underlie responses to those seven variables. The first factor contained X_1 (delivery speed), X_2 (price), X_3 (price flexibility), and X_7 (product quality; X_2 and X_7 were inversely related to X_1 and X_3); the second factor contained the two image items, X_4 (manufacturer image) and X_6 (salesforce image).

For the two-cluster solution, six of the seven variables produced statistically significantly effects (see Table 5.6). Only X_5, overall service, was not significantly different between the two clusters. In profiling customers, Cluster 1 is significantly higher on the first factor, whereas Cluster 2 has significantly higher perceptions of HATCO on the second factor—the two image variables (X_4 and X_6). The differences were much more distinctive on the first set of variables, and the image variables had substantially less delineation between the clusters, even though they were statistically significant. This should focus managerial attention on the four variables in the first set and relegate the image variables to a secondary role, although image variables are the key descriptors for the second cluster.

In the four-cluster solution, Cluster 1 split into Clusters 1 and 4, whereas Cluster 2 split into Clusters 2 and 3. Clusters 1 and 4 shared similar patterns on the first variable set (X_1, X_2, X_3, and X_7), maintaining their distinctiveness from Clusters 2 and 3. Their profiles varied, with Cluster 1 higher on X_2 and Cluster 4 higher on X_1, X_3, and X_7.

Table 5.5

Clustering Variable Profiles From the Hierarchical Cluster Analysis

Clustering Variable Profiles

	Clustering Variable Mean Values							
Cluster	X_1 Delivery speed	X_2 Price level	X_3 Price flexibility	X_4 Manufacturer image	X_5 Overall service	X_6 Salesforce image	X_7 Product quality	Cluster size
Two-cluster solution								
1	4.460	1.576	8.900	4.926	2.992	2.510	5.904	50
2	2.570	3.152	6.888	5.570	2.840	2.820	8.038	50
Four-cluster solution								
1	4.207	1.624	8.597	4.372	2.879	2.014	5.124	29
2	2.213	2.834	7.166	5.358	2.505	2.689	7.968	38
3	3.700	4.158	6.008	6.242	3.900	3.233	8.258	12
4	4.810	1.510	9.319	5.690	3.148	3.195	6.981	21

Significance Testing of Differences Between Cluster Centers

Variable	Cluster mean square	Degrees of freedom	Error mean square	Degrees of freedom	*F* value	Significance
Two-cluster solution						
X_2 Delivery speed	89.302	1	.851	98	104.95	000

Four-cluster solution								
1	4.207	1.624	8.597	4.372	2.879	2.014	5.124	29
2	2.213	2.834	7.166	5.358	2.505	2.689	7.968	38
3	3.700	4.158	6.008	6.242	3.900	3.233	8.258	12
4	4.810	1.510	9.319	5.690	3.148	3.195	6.981	21

Significance Testing of Differences Between Cluster Centers

Variable	Cluster mean square	Degrees of freedom	Error mean square	Degrees of freedom	*F* value	Significance
Two-cluster solution						
X_2 Delivery speed	89.302	1	.851	98	104.95	.000
X_2 Price level	62.094	1	.811	98	76.61	.000
X_3 Price flexibility	101.204	1	.909	98	111.30	.000
X_4 Manufacturer image	10.368	1	1.187	98	8.73	.004
X_5 Overall service	.578	1	.564	98	1.02	.314
X_6 Salesforce image	2.402	1	.576	98	4.17	.044
X_7 Product quality	113.849	1	1.377	98	82.68	.000
Four-cluster solution						
X_1 Delivery speed	37.962	3	.613	96	61.98	.000
X_2 Price level	26.082	3	.659	96	39.56	.000
X_3 Price flexibility	39.927	3	.735	96	54.34	.000
X_4 Manufacturer image	12.884	3	.917	96	14.04	.000
X_5 Overall service	6.398	3	.382	96	16.75	.000
X_6 Salesforce image	7.367	3	.383	96	19.26	.000
X_7 Product quality	52.203	3	.960	96	54.37	.000

Note. From *Multivariate Data Analysis*, by J. R. Hair, Jr., R. E. Anderson, R. L. Tatham, and W. C. Black, 1998, p. 485 Copyright 1998 by Prentice-Hall, Inc. Adapted with permission of Prentice-Hall, Inc., Upper Saddle River, NJ.

Table 5.6

Solutions of the Nonhierarchical Cluster Analysis With Initial Seed Points From the Hierarchical Results

Two-Cluster Solution

	Mean values							
Cluster	X_1 **Delivery speed**	X_2 **Price level**	X_3 **Price flexibility**	X_4 **Manufacturer image**	X_5 **Overall service**	X_6 **Salesforce image**	X_7 **Product quality**	**Cluster size**
Final cluster centers								
1	4.383	1.581	8.900	4.925	2.958	2.525	5.904	52
2	2.575	3.212	6.804	5.598	2.871	2.817	8.127	48
Statistical significance of cluster differences								
F value	87.720	86.750	133.180	9.600	0.330	3.670	96.400	
Significance	0.000	0.000	0.000	0.003	0.566	0.058	0.000	

Profiling the clusters

	Cluster			
Predictive validity	**1**	**2**	***F* value**	**Significance**
X_9 Usage level	49.212	42.729	14.789	.000
X_{10} Satisfaction	5.133	4.379	23.826	.000

Four-Cluster Solution

	Mean values							
X_1 Delivery	X_2 Price	X_3 Price	X_4 Manufacturer	X_5 Overall	X_6 Salesforce	X_7 Product	Cluster

Predictive validity	Cluster 1	Cluster 2	*F* value	Significance
X_9 Usage level	49.212	42.729	14.789	.000
X_{10} Satisfaction	5.133	4.379	23.826	.000

Four-Cluster Solution

Mean values

Cluster	X_1 Delivery speed	X_2 Price level	X_3 Price flexibility	X_4 Manufacturer image	X_5 Overall service	X_6 Salesforce image	X_7 Product quality	Cluster size
Final cluster centers								
1	4.094	1.621	8.630	4.415	2.830	2.079	5.273	33
2	2.171	2.846	7.123	5.403	2.489	2.683	8.194	35
3	3.662	4.200	5.946	6.123	3.900	3.177	7.946	13
4	4.884	1.511	9.368	5.811	3.179	3.300	7.000	19
Statistical significance of cluster differences								
F value	56.35	46.710	67.860	14.600	18.600	19.710	57.600	
Significance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

Profiling the clusters

Predictive validity	Cluster 1	Cluster 2	Cluster 3	Cluster 4	*F* value	Significance
X_9 Usage level	46.333	41.229	46.769	54.211	11.304	.000
X_{10} Satisfaction	4.839	4.134	5.038	5.642	22.212	.000

The marked difference comes on the image variables, where Cluster 1 follows the pattern seen in the two-cluster solution with lower scores, whereas Cluster 4 counters this trend by having high positive image scores. This represents the most marked difference between Clusters 1 and 4. Cluster 4 has relatively high perceptions of HATCO on all variables, whereas Cluster 1 is high only on the first set of variables.

The remaining clusters, 2 and 3, follow the pattern of having relatively high scores on the image variables and low scores elsewhere. They also differentiate themselves, however, as there is separation between them on the first set of variables, particularly X_1, X_2, and X_3. These findings reveal more in-depth profiles of types of consumers, and they can be used to develop marketing strategies directed toward each group's perceptions.

As a first validity check for stability of the cluster solution, a second nonhierarchical analysis was performed, this time allowing the procedure to randomly select the initial seed points for both cluster solutions. The results confirmed a consistency in the results for both cluster solutions. The cluster sizes were comparable for each solution, and the cluster profiles were very similar. Given the stability of the results between the specified seed points and random selection, management should feel confident that true differences do exist among customers in terms of their needs in a supplier and that the structure depicted in the cluster analysis is supported empirically.

To assess predictive validity, variables found in prior research—involving multiple regression and canonical correlation analyses—to have a relationship with clustering variables were assessed: X_9 (usage level) and X_{10} (satisfaction level). For the two-cluster solution, the univariate F ratios show that the cluster means for both variables are significantly different. The profiling process here shows that customers in Cluster 1, which rated HATCO higher on that first set of variables, had higher levels of usage and satisfaction with HATCO, as would be expected. Likewise, Cluster 2 customers had lower ratings on the first set of variables and lower ratings on the two additional variables. The four-cluster solution exhibits a similar pattern, with the clusters showing statistically significant differences on these two additional variables. Further predictive validity analyses involving additional variables not included in the clustering procedure [X_8 (firm size), X_{11} (specification buying), X_{12} (structure of procurement), X_{13} (industry type), and X_{14} (buying situation)] were undertaken and revealed that both the two-cluster and four-cluster solutions had distinctive profiles on this set of

additional variables. Only X_{13} shows no difference across the two-cluster solution, with all other variables having significant differences for the two-cluster and four-cluster solutions. For example, Cluster 1 in the two-cluster solution was characterized as primarily small firms engaged in straight rebuys with HATCO that primarily use total-value analysis in a decentralized procurement system. Similar profiles could be developed for each cluster.

The cluster analysis of the 100 HATCO customers was successful in performing a market segmentation of HATCO customers. It not only created homogeneous groupings of customers based on their perceptions of HATCO but also found that these clusters met the tests of predictive validity and distinctiveness on additional sets of variables, all necessary for achieving practical significance. The segments represent different customer perspectives of HATCO, varying in both the types of variables that are viewed most positively and the magnitude of the perceptions.

One issue unresolved to this point is the selection of the final cluster solution between the two- and four-cluster solutions. Both provide viable market segments, as they are of substantial size and do not have any small cluster sizes caused by outliers. Moreover, they meet all of the criteria for a successful market segmentation. In either instance, the clusters (market segments) represent sets of consumers with homogeneous perceptions that can be uniquely identified, thus being prime candidates for differentiated marketing programs. The researcher can use the two-cluster solution to provide a basic delineation of customers that varying in perceptions and buying behavior or use the four-cluster solution for a more complex segmentation strategy that provides a highly differentiated mix of customer perceptions and well as targeting options.

Conclusion

Cluster analysis can be a useful data reduction technique. Because its application is more of an art than a science, however, it can easily be abused or misapplied by researchers. Different interobject measures and different algorithms can and do affect the results. The selection of the final cluster solution in most cases is based on both objective and subjective considerations. The prudent researcher, therefore, considers these issues and always assesses the effect of all decisions. If the re-

searcher proceeds cautiously, however, cluster analysis can be an invaluable tool in identifying latent patterns suggesting useful groupings (clusters) of objects that are not discernible through other multivariate techniques.

Suggestions for Further Reading

Aldenderfer and Blashfield (1984) provided an excellent introduction to cluster analysis. For more in-depth coverage of the topic, Everitt's (1980) book is highly recommended.

Glossary

AGGLOMERATIVE METHODS *A hierarchical procedure* that begins with each object or observation in a separate cluster. In each subsequent step, the two *object* clusters that are most similar are combined to build a new aggregate cluster. The process is repeated until all objects are finally combined into a single cluster.

ALGORITHM A set of rules or procedures; similar to an equation.

AVERAGE LINKAGE The algorithm used in *agglomerative methods* that represents similarity as the average distance from all *objects* in one cluster to all objects in another. This approach tends to combine clusters with small variances.

CENTROID The average or mean value of the *objects* contained in the cluster on each variable, whether used in the *cluster variate* or in the validation process.

CENTROID METHOD The agglomerative *algorithm* in which similarity between clusters is measured as the distance between the two cluster *centroids*. When two clusters are combined, a new centroid is computed. Thus, cluster centroids migrate, or move, as the clusters are combined.

CITY-BLOCK APPROACH A method of calculating distances based on the sum of the absolute differences of the coordinates for the *objects*. This method assumes that the variables are uncorrelated and that unit scales are compatible.

CLUSTER CENTROID The average value of the *objects* contained in the cluster on all the variables in the *cluster variate.*

CLUSTER SEEDS The initial *centroids* or starting points for clusters. These values are selected to initiate *nonhierarchical procedures,* in which clusters are built around these prespecified points.

CLUSTER VARIATE A set of variables or characteristics representing the *objects* to be clustered and used to calculate the similarity between objects.

COMPLETE LINKAGE The agglomerative *algorithm* in which similarity is based on the maximum distance between *objects* in two clusters (the distance between the most dissimilar member of each cluster). At each stage of the agglomeration, the two clusters with the smallest maximum distance (most similar) are combined. Also referred to as *furthest-neighbor* or *diameter method.*

CRITERION VALIDITY The ability of clusters to show the expected differences on a variable not used to form the clusters. For example, if clusters were formed on performance ratings, standard marketing thought would suggest that clusters with higher performance ratings should also have higher satisfaction scores. If this were found to be so in empirical testing, then criterion validity is supported.

DENDROGRAM The graphical representation (tree graph) of the results of a *hierarchical procedure* in which each *object* is arrayed on the vertical axis, and the horizontal axis portrays the steps in the hierarchical procedure. Starting at the left side with each object represented as a separate cluster, the dendrogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster.

DIVISIVE METHOD A clustering procedure, the opposite of the *agglomerative method,* that begins with all objects in a single, large cluster that is then divided at each step into two clusters that contain the most dissimilar objects.

ENTROPY GROUP A group of objects independent of any cluster (i.e., they do not fit into any cluster) that may be considered outliers and possibly eliminated from the cluster analysis.

EUCLIDEAN DISTANCE The most commonly used measure of the similarity between two *objects.* Essentially, it is a measure of the length of a straight line drawn between two objects.

HIERARCHICAL PROCEDURES The stepwise clustering procedures involving a combination (or division) of the *objects* into clusters. The two alternative procedures are the *agglomerative* and *divisive methods.* The result is the construction of a hierarchy or treelike structure (*dendrogram*) depicting the formation of the clusters, which produces $N - 1$ cluster solutions, where N is the number of objects. For example, if the agglomerative procedure starts with five objects in separate clusters, it shows how four clusters, then three, then two, and finally one cluster is formed.

INTEROBJECT SIMILARITY The correspondence or association of two *objects* based on the variables of the *cluster variate.* Similarity can be measured in two ways. First is a measure of association, such as higher positive correlation coefficients representing greater similarity. Second, "proximity" or "closeness" between each pair of objects can assess similarity, where measures of distance or difference are used, with smaller distances or differences representing greater similarity.

MAHALANOBIS DISTANCE A standardized form of *Euclidean distance.* Scaling responses in terms of standard deviations standardizes data, and adjustments are made for intercorrelations between the variables.

MULTICOLLINEARITY The extent to which a variable can be explained by the other variables in the analysis. As multicollinearity increases, it complicates the interpretation of the variate because it is more difficult to ascertain the effect of any single variable, owing to their interrelationships.

NONHIERARCHICAL PROCEDURES Procedures in which, instead of the treelike construction process found in the hierarchical procedures, *cluster seeds* are used to group objects within a prespecified distance of the seeds. The procedures produce only a single cluster solution for a set of cluster seeds. For example, if four cluster seeds are specified, only four clusters are formed. It does not produce results for all possible number of clusters as done with a *hierarchical procedure.*

NORMALIZED DISTANCE FUNCTION A process that converts each raw data score to a standardized variate with a mean of 0 and a standard deviation of 1, as a means to remove the bias introduced by differences in scales of several variables.

OBJECT A person, product–service, firm, or any other entity that can be evaluated on a number of attributes.

OUTLIER An observation that is substantially different from other observations (i.e., has an extreme value).

OPTIMIZING PROCEDURE A *nonhierarchical procedure* that allows for the reassignment of *objects* from the originally assigned cluster to another cluster on the basis of an overall optimizing criterion.

PARALLEL THRESHOLD METHOD A *nonhierarchical procedure* that selects the *cluster seeds* simultaneously in the beginning. *Objects* within the threshold distances are assigned to the nearest seed. Threshold distances can be adjusted to include fewer or more objects in the clusters. This method is the opposite of the *sequential threshold method.*

PREDICTIVE VALIDITY See *criterion validity.*

PROFILE DIAGRAM A graphical representation of data that aids in screening for *outliers* or the interpretation of the final cluster solution. Typically, the variables of the *cluster variate* or those used for validation are listed along the horizontal axis and the value scale on the vertical axis. Separate lines depict the scores (original or standardized) for individual *objects* or cluster *centroids* in a graphic plane.

RESPONSE-STYLE EFFECT A series of systematic responses by a respondent that reflect a "bias" or consistent pattern. Examples include always responding that an *object* is excellent or poor performing across all attributes with little or no variation.

ROW-CENTERING STANDARDIZATION See *within-case standardization.*

SEQUENTIAL THRESHOLD METHOD A *nonhierarchical procedure* that begins by selecting one *cluster seed.* All *objects* within a prespecified distance are then included in that cluster. Subsequent cluster seeds are selected until all objects are grouped in a cluster.

SINGLE LINKAGE A *hierarchical procedure* where similarity is defined as the minimum distance between any *object* in one cluster and any object in another. Simply, this means the distance between the closest objects in two clusters. This procedure has the potential for creating less compact, even chainlike clusters. This differs from the cluster *centroid method* that uses some measure of all objects in the cluster.

STOPPING RULE An *algorithm* for determining the final number of clusters to be formed. With no stopping rule inherent in cluster analysis, researchers have developed several criteria and guidelines for this problem. Two classes of rules exist that are applied post hoc and

calculated by the researcher: (a) measures of similarity and (b) adapted statistical measures.

TAXONOMY An empirically derived classification of actual *objects* based on one or more characteristics. Typified by the application of cluster analysis or other grouping procedures. Can be contrasted to a *typology.*

TYPOLOGY A conceptually based classification of *objects* based on one or more characteristics. A typology does not usually attempt to group actual observations, but instead provides the theoretical foundation for the creation of a *taxonomy* that groups actual observations.

WARD'S METHOD A *hierarchical procedure* where the similarity used to join clusters is calculated as the sum of squares between the two clusters summed over all variables. It has the tendency to result in clusters of approximately equal size due to its minimization of within-group variation.

WITHIN-CASE STANDARDIZATION A method of standardization in which a respondent's responses are not compared with the overall sample but instead to their own responses. Also know as *ipsitizing,* each respondent's average response is used to standardize his or her own responses.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Thousand Oaks, CA: Sage.

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

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques.* Thousand Oaks, CA: Sage.

Campbell, R., & Johnson, C. R. (1997). Police officer's perceptions of rape: Is there consistency between state law and individual beliefs? *Journal of Interpersonal Violence, 12*(2), 255–274.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). New York: Halsted Press.

Green, P. E. (1978). *Analyzing multivariate data.* Hinsdale, IL: Holt, Rinehart & Winston.

Green, P. E., & Carroll, J. D. (1978). *Mathematical tools for applied multivariate analysis.* New York: Academic Press.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis.* Upper Saddle River, NJ: Prentice-Hall.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 15,* 225–238.

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

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179.

Mowbray, C.
Cluster
Overall, J. (1
letin, 6
Punj, G., &
suggest
Schaninger,
determ
237–25
Shephard, R.
ogy, 3,
Singh, J. (19
tailing,
Sneath, P. H.
Press.

Mowbray, C. T., Bybee, D., & Cohen, E. (1993). Describing the homeless mentally ill: Cluster analysis results. *American Journal of Community Psychology, 21*(2), 67–93.

Overall, J. (1964). Note on multivariate methods for profile analysis. *Psychological Bulletin, 61,* 195–198.

Punj, G., & Stewart, D. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20,* 134–148.

Schaninger, C. M., & Bass, W. C. (1986). Removing response-style effects in attribute-determinance ratings to identify market segments. *Journal of Business Research, 14,* 237–252.

Shephard, R. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3,* 287–315.

Singh, J. (1990). A typology of consumer dissatisfaction response styles. *Journal of Retailing, 66,* 57–99.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy.* San Francisco: Freeman Press.