UNIT - III

Cluster Analysis

Cluster Analysis

- **Cluster analysis** is a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess.
- It has been referred to as Q analysis, typology construction, classification analysis, and numerical taxonomy.
- This variety of names is due to the usage 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 relationships among the objects being clustered.
- This common dimension represents the essence of all clustering approaches—the classification of data as suggested by natural groupings of the data themselves.
- Cluster analysis is comparable to factor analysis in its objective of assessing structure.
- **Cluster analysis differs from factor analysis**, however, in that cluster analysis groups objects, whereas factor analysis is primarily concerned with grouping variables.
- Additionally, factor analysis makes the groupings based on patterns of variation (correlation) in the data whereas cluster analysis makes groupings on the basis of distance (proximity).

Cluster Analysis

- **Cluster Analysis as a Multivariate Technique**
- Cluster analysis classifies objects (e.g., respondents, products, or other entities), on a set of user selected characteristics.
- The resulting clusters should exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity.
- Thus, if the classification is successful, the objects within clusters will be close together when plotted geometrically, and different clusters will be far apart.
- The variate in cluster analysis is determined quite differently from other multivariate techniques.
- Cluster analysis is the only multivariate technique that does not estimate the variate empirically but instead uses the variate as specified by the researcher.
- **Conceptual Development with Cluster Analysis**
- Cluster analysis has been used in every research setting imaginable.
- Ranging from the derivation of taxonomies in biology for grouping all living organisms, to psychological classifications based on personality and other personal traits, to segmentation analyses of markets, cluster analysis applications have focused largely on grouping individuals.
- However, cluster analysis can classify objects other than individual people, including the market structure, analyses of the similarities and differences among new products, and performance evaluations of firms to identify groupings based on the firms' strategies or strategic orientations.

Cluster Analysis

- The more common roles cluster analysis can play in conceptual development include the following:
- **Data reduction:**
  - A researcher may be faced with a large number of observations that are meaningless unless classified into manageable groups.
  - Cluster analysis can perform this data reduction procedure objectively by reducing the information from an entire population or sample to information about specific groups.
  - 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.
- **Hypothesis generation:**
  - Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously 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 soft drinks, and the resulting clusters, if any, can be profiled for demographic similarities and differences.

Cluster Analysis

- **Necessity of Conceptual Support in Cluster Analysis :**
  - Even if cluster analysis is being used in conceptual development as just mentioned, some conceptual rationale is essential.
  - The following are the most common criticisms that must be addressed by conceptual rather than empirical support:
  - **Cluster analysis is descriptive, atheoretical, and noninferential.**
  - Cluster analysis has no statistical basis upon which to draw inferences from a sample to a population, and many contend that it is only an exploratory technique.
  - Nothing guarantees unique solutions, because the cluster membership for any number of solutions is dependent upon many elements of the procedure, and many different solutions can be obtained by varying one or more elements.
  - **Cluster analysis will always create clusters, regardless of the actual existence of any structure in the data.**
  - When using cluster analysis, the researcher is making an assumption of some structure among the objects.
  - The researcher should always remember that just because clusters can be found does not validate their existence.
  - Only with strong conceptual support and then validation are the clusters potentially meaningful and relevant.

Cluster Analysis

- **Necessity of Conceptual Support in Cluster Analysis :**
  - **The cluster solution is not generalizable because it is totally dependent upon the variables used as the basis for the similarity measure.**
  - This criticism can be made against any statistical technique, but cluster analysis is generally considered more dependent on the measures used to characterize the objects than other multivariate techniques.
  - With the cluster variate completely specified by the researcher, the addition of spurious variables or the deletion of relevant variables can have a substantial impact on the resulting solution.
  - As a result, the researcher must be especially cognizant of the variables used in the analysis, ensuring that they have strong conceptual support.
  - Thus, in any use of cluster analysis the researcher must take particular care in ensuring that strong conceptual support predates the application of the technique.

Cluster Analysis

- **HOW DOES CLUSTER ANALYSIS WORK?**
  - Cluster analysis performs a task innate to all individuals—pattern recognition and grouping.
  - The human ability to process even slight differences in innumerable characteristics is a cognitive process inherent in human beings that is not easily matched with all of our technological advances.
  - Take for example the task of analyzing and grouping human faces. Even from birth, individuals can quickly identify slight differences in facial expressions and group different faces in homogeneous groups while considering hundreds of facial characteristics.
  - Yet we still struggle with facial recognition programs to accomplish the same task. The process of identifying natural groupings is one that can become quite complex rather quickly.
  - To demonstrate how cluster analysis operates, we examine a simple example that illustrates some of the key issues: measuring similarity, forming clusters, and deciding on the number of clusters that best represent structure.

Cluster Analysis

- **HOW DOES CLUSTER ANALYSIS WORK?**
  - **A Simple Example**
  - The nature of cluster analysis and the basic decisions on the part of the researcher will be illustrated by a simple example involving identification of customer segments in a retail setting.
  - Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores.
  - A small sample of seven respondents is selected as a pilot test of how cluster analysis is applied.
  - Two measures of loyalty—V1 (store loyalty) and V2 (brand loyalty)— were measured for each respondent on a 0–10 scale.
  - The values for each of the seven respondents are shown in Figure 1, along with a scatter diagram depicting each observation on the two variables.
  - The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups.

Cluster Analysis

- **HOW DOES CLUSTER ANALYSIS WORK?**
  - To accomplish this task, we must address three basic questions:
  - **1. How do we measure similarity?**
  - We require a method of simultaneously comparing observations on the two clustering variables (V1 and V2).
  - Several methods are possible, including the correlation between objects or perhaps a measure of their proximity in two-dimensional space such that the distance between observations indicates similarity.
  - **2. How do we form clusters?**
  - No matter how similarity is measured, the procedure must group those observations that are most similar into a cluster, thereby determining the cluster group membership of each observation for each set of clusters formed.
  - **3. How many groups do we form?**
  - The final task is to select one set of clusters as the final solution.
  - In doing so, the researcher faces a trade-off: fewer clusters and less homogeneity within clusters versus a larger number of clusters and more within-group homogeneity.
  - Yet as the number of clusters decreases, the heterogeneity within the clusters necessarily increases.
  - Thus, a balance must be made between defining the most basic structure (fewer clusters) that still achieves an acceptable level of heterogeneity between the clusters.

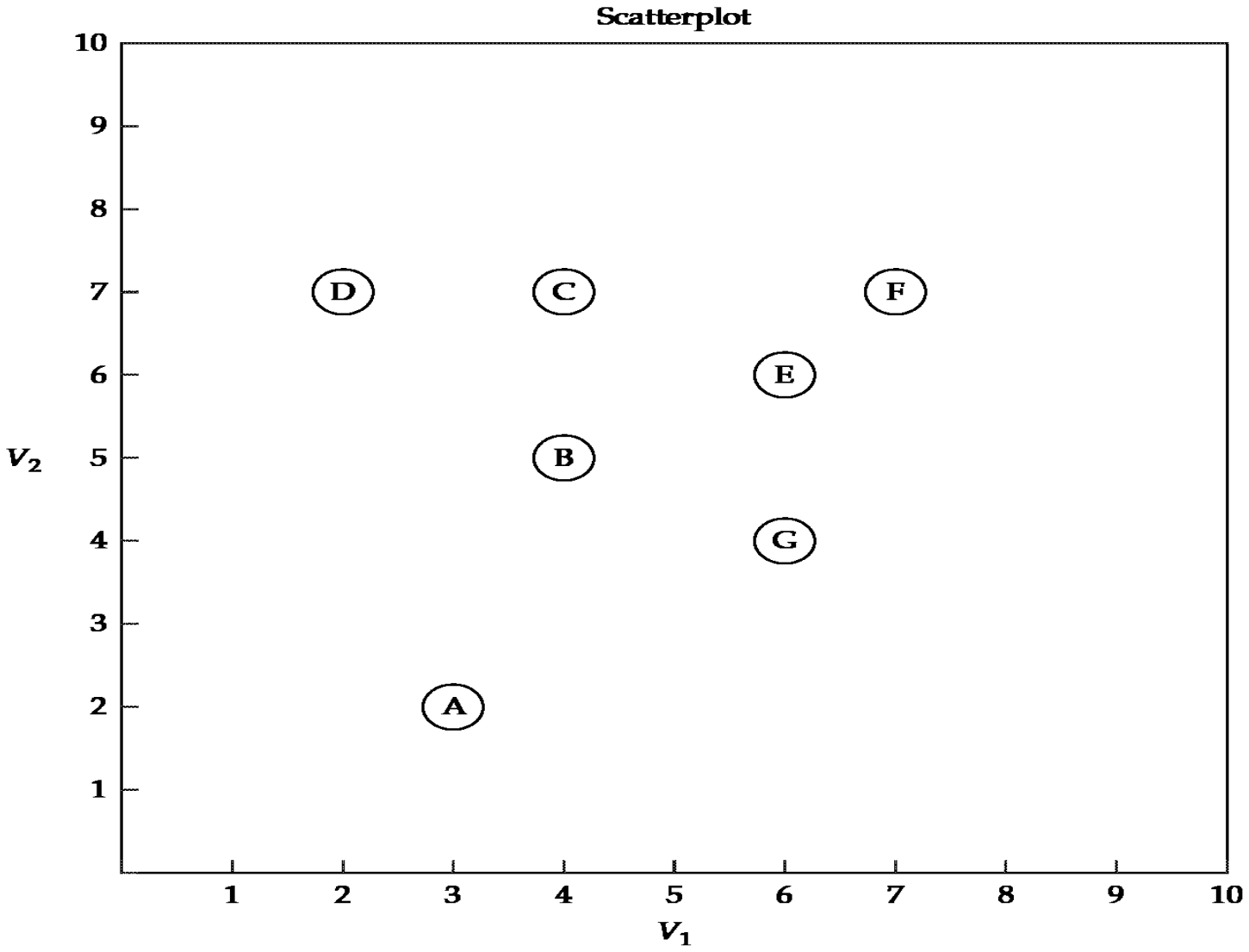
Cluster Analysis

Cluster Analysis

Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Cluster Analysis



Cluster Analysis

- **MEASURING SIMILARITY :**
  - The first task is developing some measure of similarity between each object to be used in the clustering process.
  - Similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis.
  - In a way, similarity measures are more descriptively dissimilarity measures in that smaller numbers represent greater similarity and larger numbers represent less similarity.
  - Similarity must be determined between each of the seven observations (respondents A–G) to enable each observation to be compared to each other.
  - In this example, similarity will be measured according to the Euclidean (straight-line) distance between each pair of observations (see Table 1) based on the two characteristics (V1 and V2).
  - In this two-dimensional case (where each characteristic forms one axis of the graph) we can view distance as the proximity of each point to the others.
  - 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), and A and F are the most dissimilar (6.403).

Cluster Analysis

- **MEASURING SIMILARITY :**

TABLE 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	—

Cluster Analysis

- **FORMING CLUSTERS :**
  - With similarity measures calculated, we now move to forming clusters based on the similarity measure of each observation.
  - Typically we form a number of cluster solutions (a two-cluster solution, a three-cluster solution, etc.).
  - Once clusters are formed, we then select the final cluster solution from the set of possible solutions.
  - First we will discuss how clusters are formed and then examine the process for selecting a final cluster solution.
  - Having calculated the similarity measure, we must develop a procedure for forming clusters.
  - We use this simple rule:
  - Identify the two most similar (closest) observations not already in the same cluster and combine them.
  - We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own “cluster” and then combining two clusters at a time until all observations are in a single cluster.
  - This process 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 combining existing clusters.

Cluster Analysis

- **FORMING CLUSTERS :**

TABLE 2 Agglomerative Hierarchical Clustering Process

	AGGLOMERATION PROCESS		CLUSTER SOLUTION		
Step	*Minimum Distance Between Unclustered Observations*[a]	*Observation Pair*	*Cluster Membership*	*Number 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

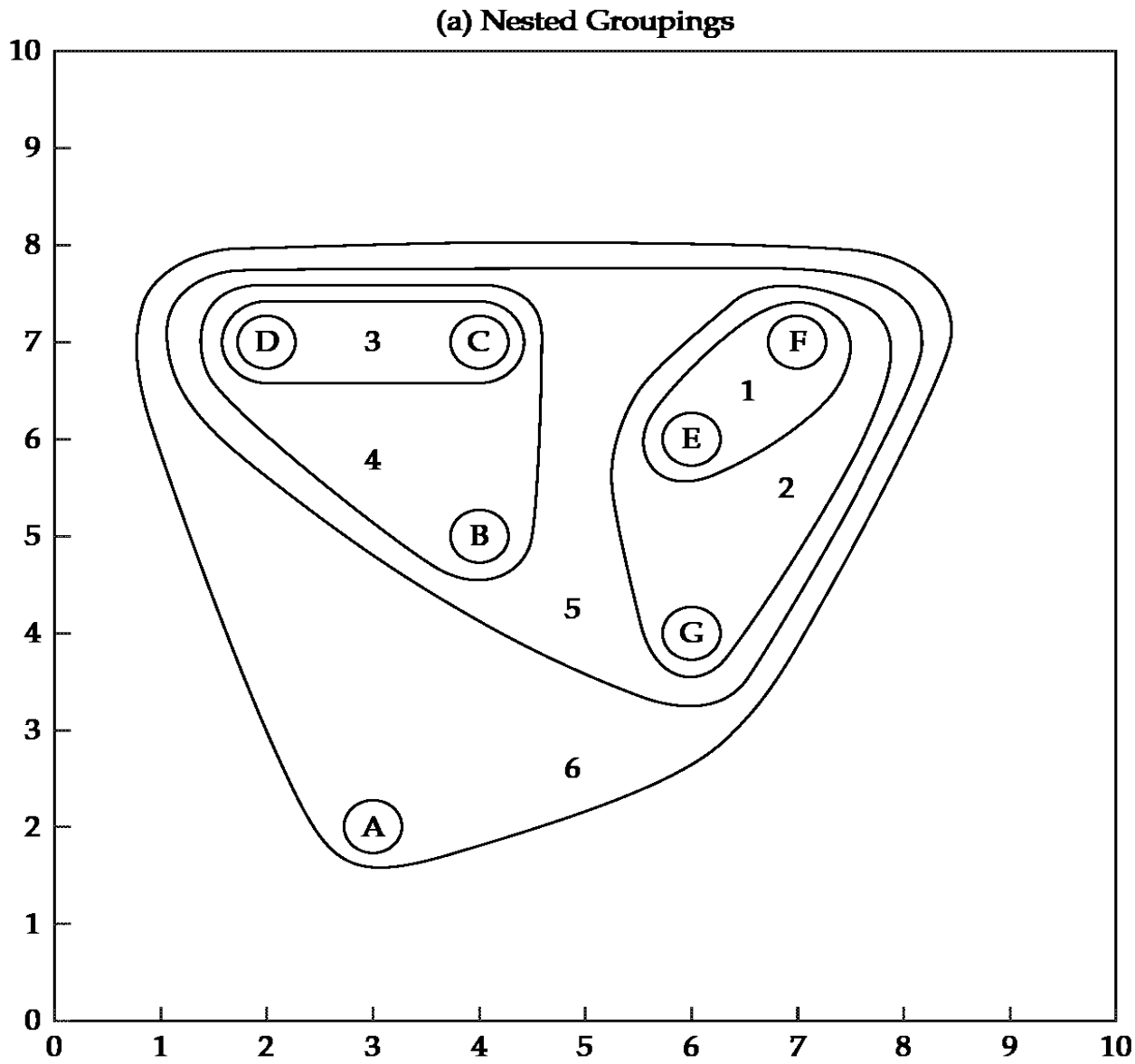
Cluster Analysis

- **FORMING CLUSTERS :**
  - The six-step clustering process is described here:
  - **Step 1:** Identify the two closest observations (E and F) and combine them into a cluster, moving from seven to six clusters.
  - **Step 2:** Find the next closest pairs of observations. In this case, three pairs have the same distance of 2.000 (E-G, C-D, and B-C). For our purposes, choose the observations E-G. G is a single-member cluster, but E was combined in the prior step with F. So, the cluster formed at this stage now has three members: G, E, and F.
  - **Step 3:** Combine the single-member clusters of C and D so that we now have four clusters.
  - **Step 4:** Combine B with the two-member cluster C-D that was formed in step 3. At this point, we now have three clusters: cluster 1 (A), cluster 2 (B, C, and D), and cluster 3 (E, F, and G).
  - **Step 5:** Combine the two three-member clusters into a single six-member cluster. The next smallest distance is 2.236 for three pairs of observations (E-B, B-G, and C-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, and D versus E, F, and G).
  - **Step 6**: Combine observation A with the remaining cluster (six observations) into a single cluster at a distance of 3.162. You will note that distances smaller or equal to 3.162 are not used because they are between members of the same cluster.

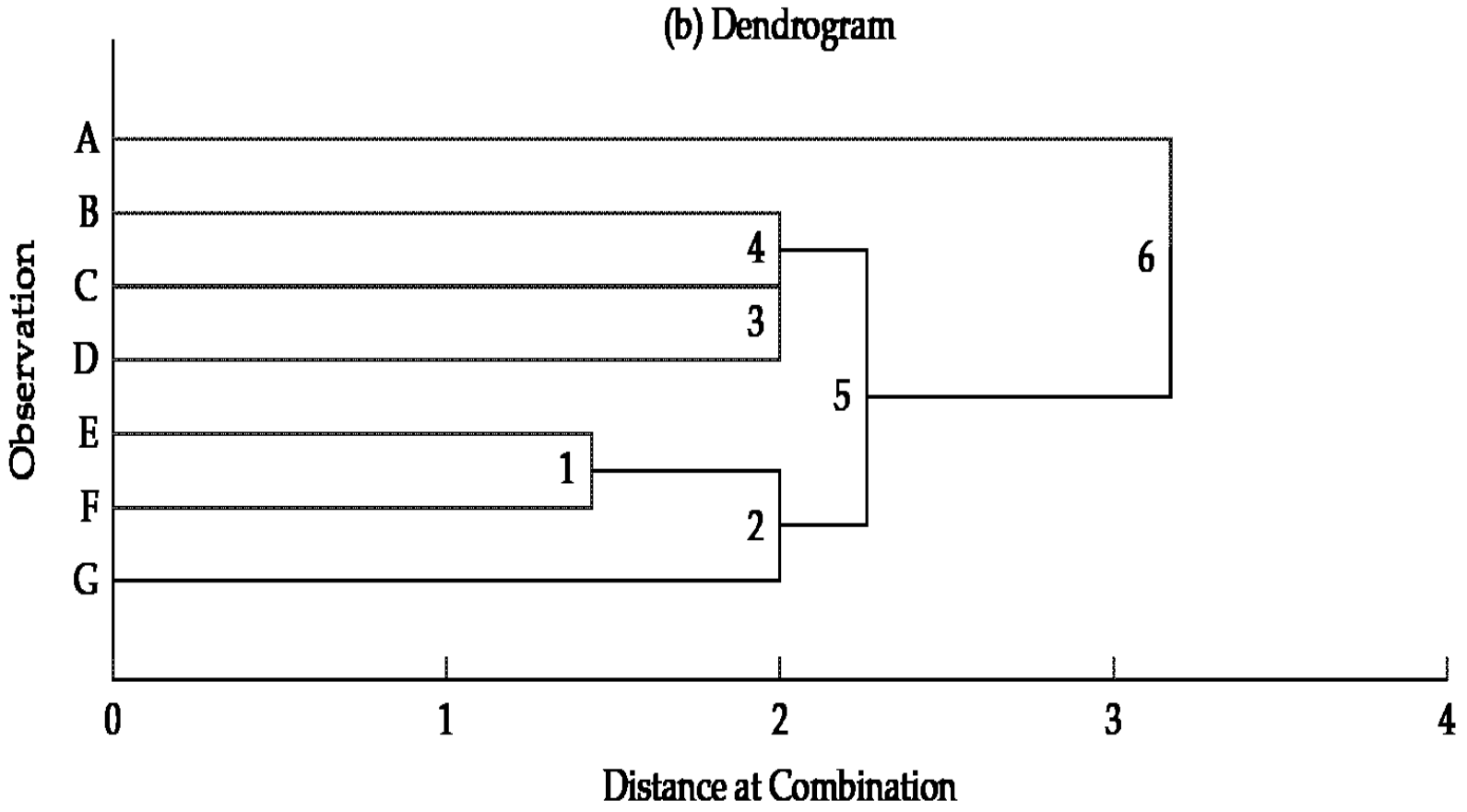
Cluster Analysis

- **FORMING CLUSTERS :**
  - The hierarchical clustering process can be portrayed graphically in several ways. Figure 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 2a).
  - This process, however, can represent the proximity of the observations for only two or three clustering variables in the scatterplot or three dimensional graph.
  - A more common approach is a **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 particularly 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.

Cluster Analysis



Cluster Analysis



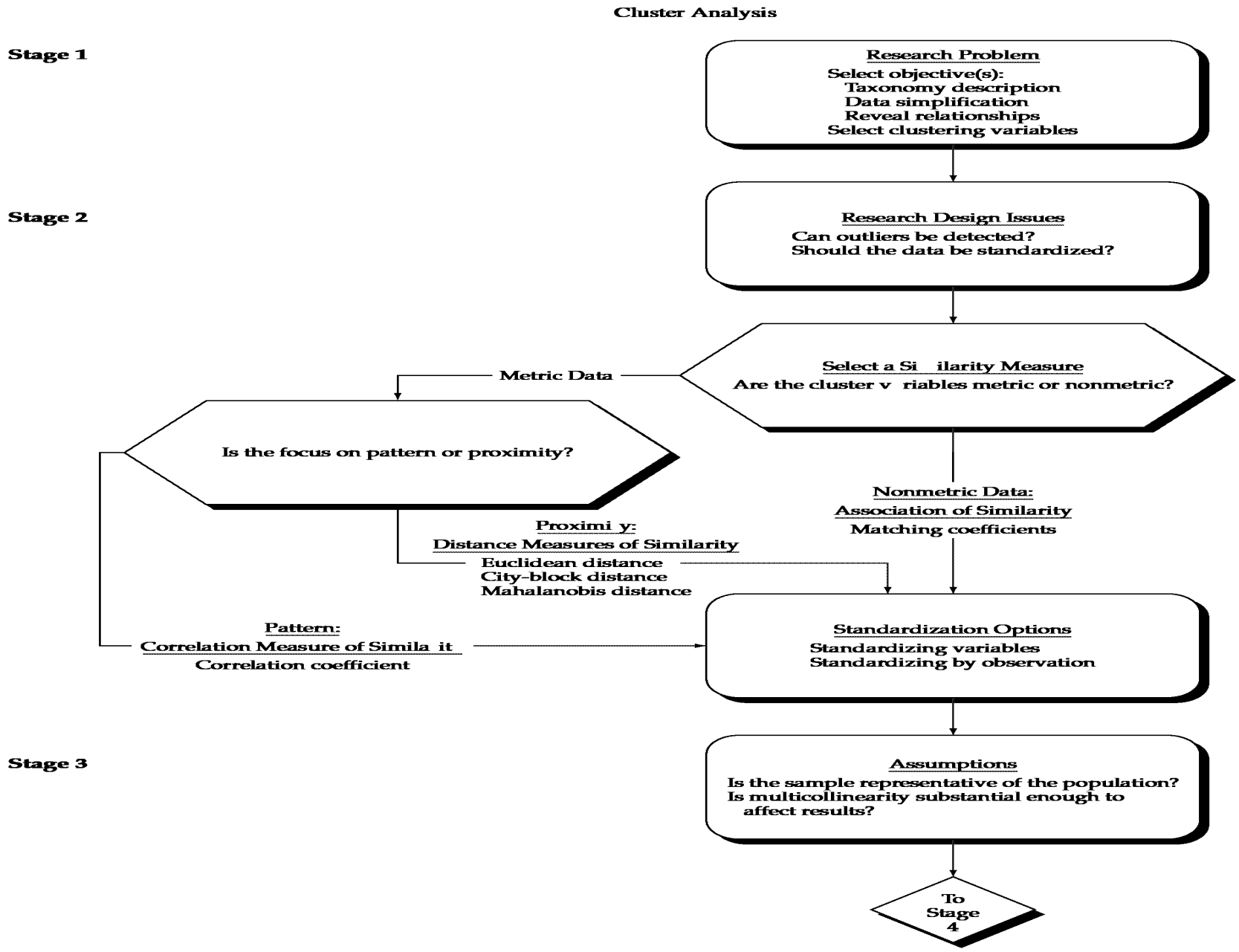
Cluster Analysis

- **DETERMINING THE NUMBER OF CLUSTERS IN THE FINAL SOLUTION :**
  - A hierarchical method results in a number of cluster solutions—in this case starting with a seven-cluster solution and ending in a one-cluster solution.
  - If all observations are treated as their own unique cluster, no data reduction has taken place and no true segments have been found.
  - The goal is identifying segments by combining observations, but at the same time introducing only small amounts of heterogeneity.
  - **Measuring Heterogeneity :**
  - Any measure of heterogeneity of a cluster solution should represent the overall diversity among observations in all clusters.
  - In the initial solution of an agglomerative approach where all observations are in separate clusters, no heterogeneity exists.
  - As observations are combined to form clusters, heterogeneity increases.
  - The measure of heterogeneity thus should start with a value of zero and increase to show the level of heterogeneity as clusters are combined.

Cluster Analysis

- **DETERMINING THE NUMBER OF CLUSTERS IN THE FINAL SOLUTION :**
  - **Measuring Heterogeneity :**
  - In this example, we use a simple measure of heterogeneity:
  - the average of all distances between observations within clusters (see Table 2).
  - As already described, the measure should increase as clusters are combined:
  - • In the initial solution with seven clusters, our overall similarity measure is 0—no observation is paired with another.
  - • **Six clusters:** The overall similarity is the distance between the two observations (1.414) joined in step 1.
  - • **Five clusters:** Step 2 forms a three-member cluster (E, F, and G), so that the overall similarity measure 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.
  - • **Four clusters:** In the next step a new two-member cluster is formed with a distance of 2.000, which causes the overall average to fall slightly to 2.144.
  - • **Three, two, and one clusters:** The final three steps form new clusters in this manner until a single-cluster solution is formed (step 6), in which the average of all distances in the distance matrix is 3.420.

Cluster Analysis Decision Process



Cluster Analysis Decision Process

- Cluster analysis, like the other multivariate techniques discussed earlier, 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 following:
  - • Partitioning the data set to form clusters and selecting a cluster solution
  - • Interpreting the clusters to understand the characteristics of each cluster and develop a name or label that appropriately defines its nature
  - • Validating the results of the final 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
- **Stage 1: Objectives of Cluster Analysis**
  - 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 fulfilling this basic objective, the researcher must address two key issues:
  - the research questions being addressed in this analysis and
  - the variables used to characterize objects in the clustering process.

Cluster Analysis Decision Process

- **RESEARCH QUESTIONS IN CLUSTER ANALYSIS**
- In forming homogeneous groups, cluster analysis may address any combination of three basic research questions:
- **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.
  - Cluster analysis can also generate hypotheses related to the structure of the objects.
  - Finally, although viewed principally as an exploratory technique, cluster analysis can be used for confirmatory purposes.
  - In such cases, a proposed typology (theoretically based classification) can be compared to that derived from the cluster analysis.
- **2. Data simplification.**
  - By defining structure among the observations, cluster analysis also develops a simplified perspective by grouping observations 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 clusters and profiled by their general characteristics

Cluster Analysis Decision Process

- **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 typically is not possible with the individual observations.
  - Whether analyses such as discriminant analysis are used to empirically identify relationships, or the groups are examined by more qualitative methods, the simplified structure from cluster analysis often identifies relationships or similarities and differences not previously revealed.
- **SELECTION OF CLUSTERING VARIABLES**
  - The objectives of cluster analysis cannot be separated from the selection of variables used to characterize the objects being clustered.
  - Whether the objective is exploratory or confirmatory, the researcher effectively constrains the possible results by the variables selected for use.
  - The derived clusters reflect the inherent structure of the data and are defined only by the variables.
  - Thus, selecting the variables to be included in the cluster variate must be done with regard to theoretical and conceptual as well as practical considerations.

Cluster Analysis Decision Process

- **Conceptual Considerations**
  - Any application of cluster analysis must have some rationale upon 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
  - (1) characterize the objects being clustered
  - (2) relate specifically to the objectives of the cluster analysis.
  - The cluster analysis technique has no means of differentiating relevant from irrelevant variables and derives the most consistent, yet distinct, groups of objects across all variables.
  - Thus, one should never include variables indiscriminately.
  - Instead, carefully choose the variables with the research objective as the criterion for selection.
- **Practical Considerations**
  - Cluster analysis can be affected dramatically by the inclusion of only one or two inappropriate or undifferentiated variables.
  - We always encouraged to examine the results and to eliminate the variables that are not distinctive (i.e., that do not differ significantly) across the derived clusters.
  - This procedure enables the cluster techniques to maximally define clusters based only on those variables exhibiting differences across the objects.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - With the objectives defined and variables selected, we must address four questions before starting the partitioning process:
  - 1. Is the sample size adequate?
  - 2. Can outliers be detected and, if so, should they be deleted?
  - 3. How should object similarity be measured?
  - 4. Should the data be standardized?
  - Many different approaches can be used to answer these questions. However, none of them 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 as much an art as a science.
  - **SAMPLE SIZE**
  - The issue of sample size in cluster analysis does not relate to any statistical inference issues (i.e., statistical power).
  - Instead the sample size must be large enough to provide sufficient representation of small groups within the population and represent the underlying structure.
  - Small groups will naturally appear as small numbers of observations, particularly when the sample size is small.
  - For example, when a sample contains only 100 or fewer observations, groups that actually make up 10 percent of the population may be represented by only one or two observations due to the sampling process.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - In such instances the distinction between outliers and representatives of a small group is much harder to make.
  - Larger samples increase the chance that small groups will be represented by enough cases to make their presence more easily identified.
  - As a result, the researcher should ensure the sample size is sufficiently large to adequately represent all of the relevant groups of the population.
  - Obviously, if the analysis objectives require identification of small groups within the population, the researcher should strive for larger sample sizes.
  - If we are interested only in larger groups (e.g., major segments for promotional campaigns), however, then the distinction between an outlier and a representative of a small group is less important and they can both be handled in a similar manner.

Cluster Analysis Decision Process

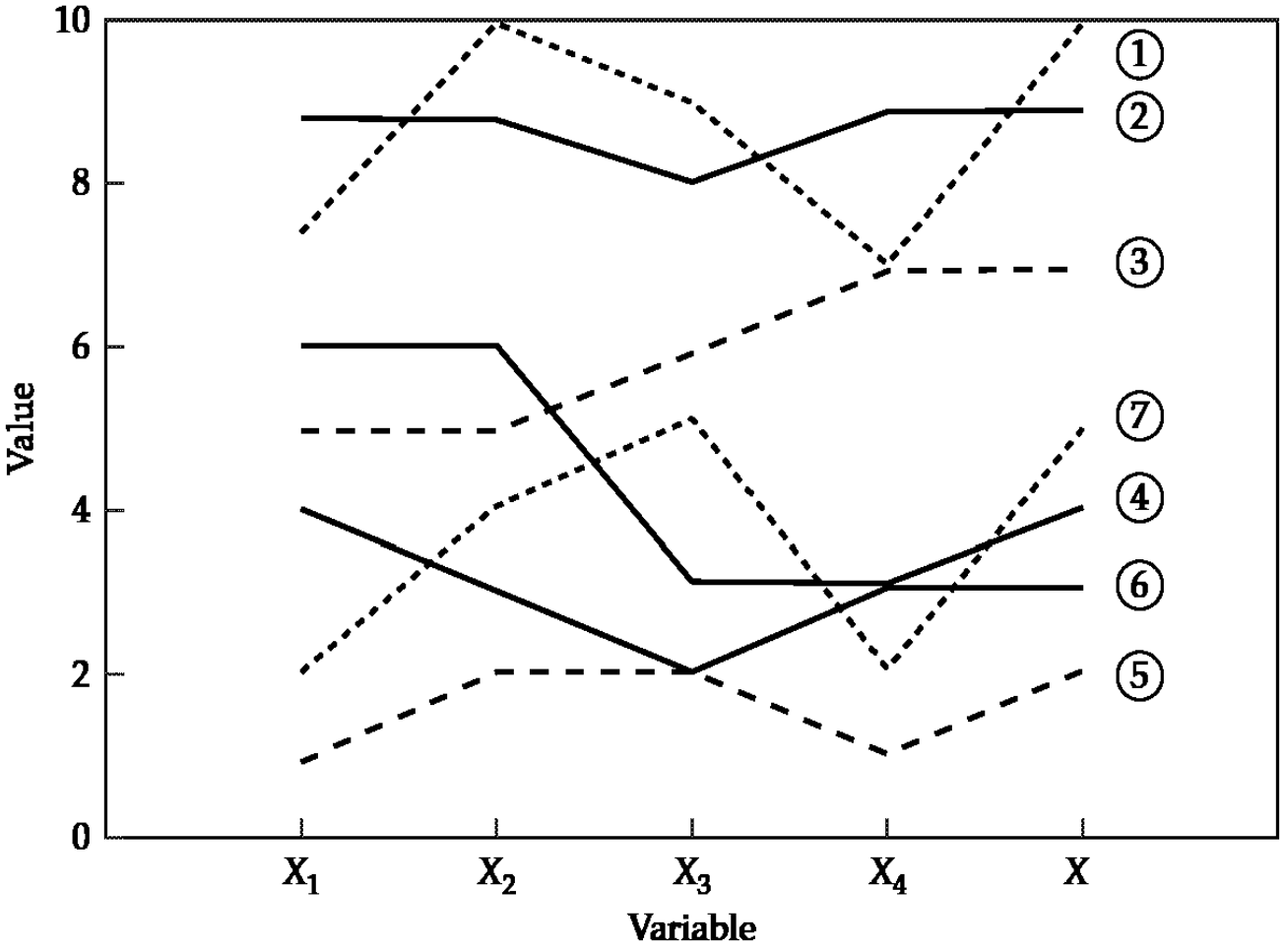
- **Stage 2: Research Design in Cluster Analysis**
  - **DETECTING OUTLIERS**
  - In its search for structure, cluster analysis is sensitive to the inclusion of irrelevant variables. But cluster analysis is also sensitive to outliers (objects different from all others).
  - Outliers can represent either:
  - Truly abnormal observations that are not representative of the general population.
    - In this case they distort the actual structure & make the derived clusters unrepresentative of the population structure ⟶ DELETE IT.
  - Representative observations of small or insignificant groups of objects within the population
    - In the second case, the outlier is removed so that the resulting clusters more accurately represent the relevant segments in the populations. ⟶ DELETE IT.
  - An undersampling of actual group(s) in the population that causes poor representation of the group(s) in the sample
    - However, in the third case the outliers should be included in the cluster solutions, even if they are underrepresented in the sample, because they represent valid and relevant groups. ⟶ KEEP IT.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **DETECTING OUTLIERS**
  - For this reason, a preliminary screening for outliers is always necessary.
  - **Graphical Approaches**
  - One of the simplest ways to screen data for outliers is to prepare a graphic profile diagram, listing 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 respondents that have very different profiles from the more typical respondents.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **PROFILE DIAGRAM :**



Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **DETECTING OUTLIERS**
  - **Empirical Approaches**
  - Although quite simple, the graphical procedures become difficult with a large number of objects and even more difficult as the number of variables increases.
  - Moreover, detecting outliers must extend beyond a univariate approach, because outliers also may be defined in a multivariate sense as having unique profiles across an entire set of variables that distinguish them from all of the other observations.
  - As a result, an empirical measure is needed to facilitate comparisons across objects.
  - Another approach is to identify outliers through the measures of similarity.
  - The most obvious examples of outliers are single observations that are the most dissimilar to the other observations.
  - Before the analysis, the similarities of all observations can be compared to the overall group centroid (typical respondent).
  - Isolated observations showing great dissimilarity can be dropped. Clustering patterns can also be observed once the cluster program has been run.
  - Thus, emphasis should be placed on identifying outliers before the analysis begins. Since, they can affect the result of how clusters are formed.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **MEASURING SIMILARITY**
  - The concept of similarity is fundamental to cluster analysis.
  - Interobject similarity is an empirical measure of correspondence, or resemblance, between objects to be clustered.
  - Comparing the two interdependence techniques (factor analysis and cluster analysis) will demonstrate how similarity works to define structure in both instances.
  - In factor analysis :
    - the correlation matrix between all pairs of variables was used to group variables into factors.
    - The correlation coefficient represented the similarity of each variable to another variable when viewed across all observations.
    - Thus, factor analysis grouped together variables that had high correlations among themselves.
  - A comparable process occurs in cluster analysis :
    - Here, the similarity measure is calculated for all pairs of objects, with similarity based on the profile of each observation across the characteristics specified by the researcher.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - In this way, any object can be compared to any other object through the similarity measure, just as we used correlations between variables in factor analysis.
  - 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:
  - correlational measures, distance measures, and association measures.
- Both the correlational and distance measures require metric data, whereas the association measures are for nonmetric 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 data 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 (the correspondence of patterns across the characteristics) and low correlations denote a lack of it.
- This procedure is also followed in the application of Q-type factor analysis.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**

TABLE 3 Calculating Correlational and Distance Measures of Similarity

Original Data

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	1.00						
2	−.147	1.00					
3	.000	.000	1.00				
4	.087	.516	−.824	1.00			
5	.963	−.408	.000	− 060	1.00		
6	−.466	.791	−.354	699	−.645	1.00	
7	.891	−.516	.165	− 239	.963	−.699	1.00

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	10.25	10.20	6.00	*nc*			
5	15.78	16.19	10.10	7.07	*nc*		
6	13 1	13.00	7.28	3.87	3.87	*nc*	
7	11 27	12.16	6.32	5.10	4.90	4.36	*nc*

nc = distances ot calculated.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - A correlational measure of similarity does not look at the observed mean value, or magnitude, but instead at the patterns of movement seen as one traces the data for each case over the variables measured; in other words, the similarity in the profiles for each case.
  - In Table 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 correlations.
  - 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.
  - Correlations represent patterns across the variables rather than the magnitudes, which are comparable to a Q-type factor analysis.
  - Correlational measures are rarely used because emphasis in most applications of cluster analysis is on the magnitudes of the objects, not the patterns of values.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Distance Measures**
  - Even though 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.
  - Instead, the most commonly used measures of similarity are distance measures.
  - These distance measures represent similarity as the proximity of observations to one another across the variables in the cluster variate.
  - 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 using distance measures was shown in our 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.
  - Even though proximity may seem to be a simple concept, several distance measures are available, each with specific characteristics.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Distance Measures**
  - **Euclidean distance**
  - Euclidean distance is the most commonly recognized measure of distance, many times referred to as straight-line distance.
  - Suppose that two points in two dimensions have coordinates (X1, Y1) and (X2, Y2), 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.

$$D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$

  - **Squared (or absolute) Euclidean distance**
  - Squared (or absolute) Euclidean distance is the sum of the squared differences without taking the square root.
  - The squared Euclidean distance has the advantage of not having to take the square root, which speeds computations markedly.
  - It is the recommended distance measure for the centroid and Ward's methods of clustering.

$$D_2(i,j) = A^2 + B^2 = (X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2.$$

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Distance Measures**
  - **City-block (Manhattan) distance**
  - City-block (Manhattan) distance is not based on Euclidean distance.
  - Instead, it uses the sum of the absolute differences of the variables (i.e., the two sides of a right triangle rather than the hypotenuse).
  - This procedure is the simplest to calculate, but may lead to invalid clusters if the clustering variables are highly correlated.

$$D_3(i,j) = |A| + |B| = |X_{1i} - X_{1j}| + |X_{2i} - X_{2j}|.$$

  - **Mahalanobis distance**
  - Mahalanobis distance (D2) is a generalized distance measure that accounts for the correlations among variables in a way that weights each variable equally.
  - It also relies on standardized variables and will be discussed in more detail in a following section.
  - Although desirable in many situations, it is not available as a proximity measure in either SAS or SPSS.
  - Other distance measures are available in many clustering programs.
  - We are encouraged to explore alternative cluster solutions obtained when using different distance measures in an effort to best represent the underlying data patterns.
  - Although these distance measures are said to represent similarity, in a very real sense they better represent dissimilarity, because higher values typically mean relatively less similarity.
  - Greater distance means observations are less similar.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Comparison to Correlational Measures**
  - Distance measures focus on the magnitude of the values and portray as similar the objects that are close together, even if they have different patterns across the variables.
  - In contrast, correlation measures focus on the patterns across the variables and do not consider the magnitude of the differences between objects.
  - Let us look at our seven observations to see how these approaches differ.
  - Table 3 contains the values for the seven observations on the five variables (X1 to X5), along with both distance and correlation measures of similarity.
  - Cluster solutions using either similarity measure seem to indicate three clusters, but the membership in each cluster is quite different.
  - With smaller distances representing greater similarity, we see that cases 1 and 2 form one group (distance of 3.32), and cases 4, 5, 6, and 7 (distances ranging from 3.87 to 7.07) make up another group.
  - The distinctiveness of these two groups from each other is shown in that the smallest distance between the two clusters is 10.20.
  - These two clusters represent observations with higher versus lower values.
  - A third group, consisting of only case 3, differs from the other two groups because it has values that are both low and high.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Comparison to Correlational Measures**
  - Using the correlation as the measure of similarity, three clusters also emerge. First, cases 1, 5, and 7 are all highly correlated (.891 to .963), as are cases 2, 4, and 6 (.516 to .791).
  - Moreover, the correlations between clusters are generally close to zero or even negative.
  - Finally, case 3 is again distinct from the other two clusters and forms a single-member cluster.
  - **Which Distance Measure Is Best?**
  - In attempting to select a particular distance measure, the researcher should remember the following caveats:
  - Different distance measures or a change in 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.
  - When the variables are correlated (either positively or negatively), the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for correlations and weights all variables equally.
  - Alternatively, we may wish to avoid using highly redundant variables as input to cluster analysis.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **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 agreement occurred (both respondents said yes to a question or both said no) across the set of questions.
  - Extensions of this simple matching coefficient have been developed to accommodate multicategory nominal variables and even ordinal measures.
  - **Selecting a Similarity Measure**
  - Although three different forms of similarity measures are available, the most frequently used and preferred form is the distance measure for several reasons.
  - First, the distance measure best represents the concept of proximity, which is fundamental to cluster analysis. Correlational measures, although having widespread application in other techniques, represent patterns rather than proximity.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Selecting a Similarity Measure**
  - Second, cluster analysis is typically associated with characteristics measured by metric variables.
  - In some applications, nonmetric characteristics dominate, but most often the characteristics are represented by metric measures making distance again the preferred measure.
  - **STANDARDIZING THE DATA**
  - Clustering variables that are not all of the same scale should be standardized whenever necessary to avoid instances where a variable's influence on the cluster solution is greater than it should be.
  - We will now examine several approaches to standardization:
  - **Standardizing the 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 option can be found in all computer programs and many times is even directly included in the cluster analysis procedure.
  - The process converts each raw data score into a standardized value with a mean of 0 and a standard deviation of 1, and in turn, eliminates the bias introduced by the differences in the scales of the several attributes or variables used in the analysis.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Standardizing the Variables**
  - There are two primary benefits from standardization.
  - First, it is much easier to compare between 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 are below.
  - The magnitude represents the number of standard deviations the original value is from the mean.
  - Second, no difference occurs in the standardized values when only the scale changes.
  - For example, when we standardize a measure of time duration, the standardized values are the same whether measured in minutes or seconds.
  - Thus, using standardized variables truly eliminates the effects due to scale differences not only across variables, but for the same variable as well.
  - The need for standardization is minimized when all of the variables are measured on the same response scale (e.g., a series of attitudinal questions), but becomes quite important whenever variables using quite different measurement scales are included in the cluster variate.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Using a Standardized Distance Measure**
  - A measure of Euclidean distance that directly incorporates a standardization procedure is the Mahalanobis distance (D2).
  - The Mahalanobis approach not only performs a standardization process on the data by scaling in terms of the standard deviations but it also sums the pooled within-group variance–covariance, which adjusts for correlations among the variables.
  - Highly correlated sets of variables in cluster analysis can implicitly overweight one set of variables in the clustering procedures.
  - In short, the Mahalanobis generalized distance procedure computes a distance measure between objects comparable to R2 in regression analysis.
  - Although many situations are appropriate for use of the Mahalanobis distance, not all programs include it as a measure of similarity.
  - In such cases, we usually selects the squared Euclidean distance.

Cluster Analysis Decision Process

- **Stage 2: Research Design in Cluster Analysis**
  - **Standardizing by Observation**
  - Suppose we collected a number of ratings on a 10-point scale of the importance for several attributes used in purchase decisions for a product.
  - We could apply cluster analysis and obtain clusters, but one distinct possibility is that what we would get are clusters of people who said everything was important, some who said everything had little importance, and perhaps some clusters in between.
  - What we are seeing are patterns of responses specific to an individual.
  - These patterns may reflect a specific way of responding to a set of questions, such as yes or no.
  - If we want to identify groups according to their response style and even control for these patterns, then the typical standardization through calculating Z scores is not appropriate.
  - 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 quite effective in removing response style effects and is especially suited to many forms of attitudinal data.

Cluster Analysis Decision Process

- **Stage 3: Assumptions in Cluster Analysis**
  - Cluster analysis is not a statistical inference technique in which parameters from a sample are assessed as representing a population.
  - Instead, cluster analysis is a method for quantifying the structural characteristics of a set of observations.
  - As such, it has strong mathematical properties but not statistical foundations.
  - The requirements of normality, linearity, and homoscedasticity that were so important in other techniques really have little bearing on cluster analysis.
  - We must focus, however, on two other critical issues: representativeness of the sample and multicollinearity among variables in the cluster variate.
  - **REPRESENTATIVENESS OF THE SAMPLE**
  - Rarely does we have a census of the population to use in the cluster analysis.
  - Usually, a sample of cases is obtained and the clusters are derived in the hope that they represent the structure of the population.
  - We 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.
  - We must realize that cluster analysis is only as good as the representativeness of the sample.
  - Therefore, all efforts should be made to ensure that the sample is representative and the results are generalizable to the population of interest.

Cluster Analysis Decision Process

- **Stage 3: Assumptions in Cluster Analysis**
  - **MULTICOLLINEARITY**
  - Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.
  - **Reasons why multicollinearity occurs :**
  - It is caused by an inaccurate use of dummy variables.
  - It is caused by the inclusion of a variable which is computed from other variables in the data set.
  - Multicollinearity can also result from the repetition of the same kind of variable.
  - Generally occurs when the variables are highly correlated to each other.

Cluster Analysis Decision Process

- **Stage 3: Assumptions in Cluster Analysis**
  - **IMPACT OF MULTICOLLINEARITY**
  - Multicollinearity was an issue in other multivariate techniques because of the difficulty in discerning the true impact of multicollinear variables.
  - In cluster analysis the effect is different because multicollinearity is actually a form of implicit weighting.
  - 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, either reduce the variables to equal numbers in each set or use a distance measure that takes multicollinearity into account.
  - Another possible solution involves factoring the variables prior to clustering and either selecting one cluster variable from each factor or using the resulting factor scores as cluster variables.
  - Recall that principal components or varimax rotated factors are uncorrelated. In this way, the research can take a proactive approach to dealing with multicollinearity.

UNIT - III

Cluster Analysis

Cluster Analysis Decision Process

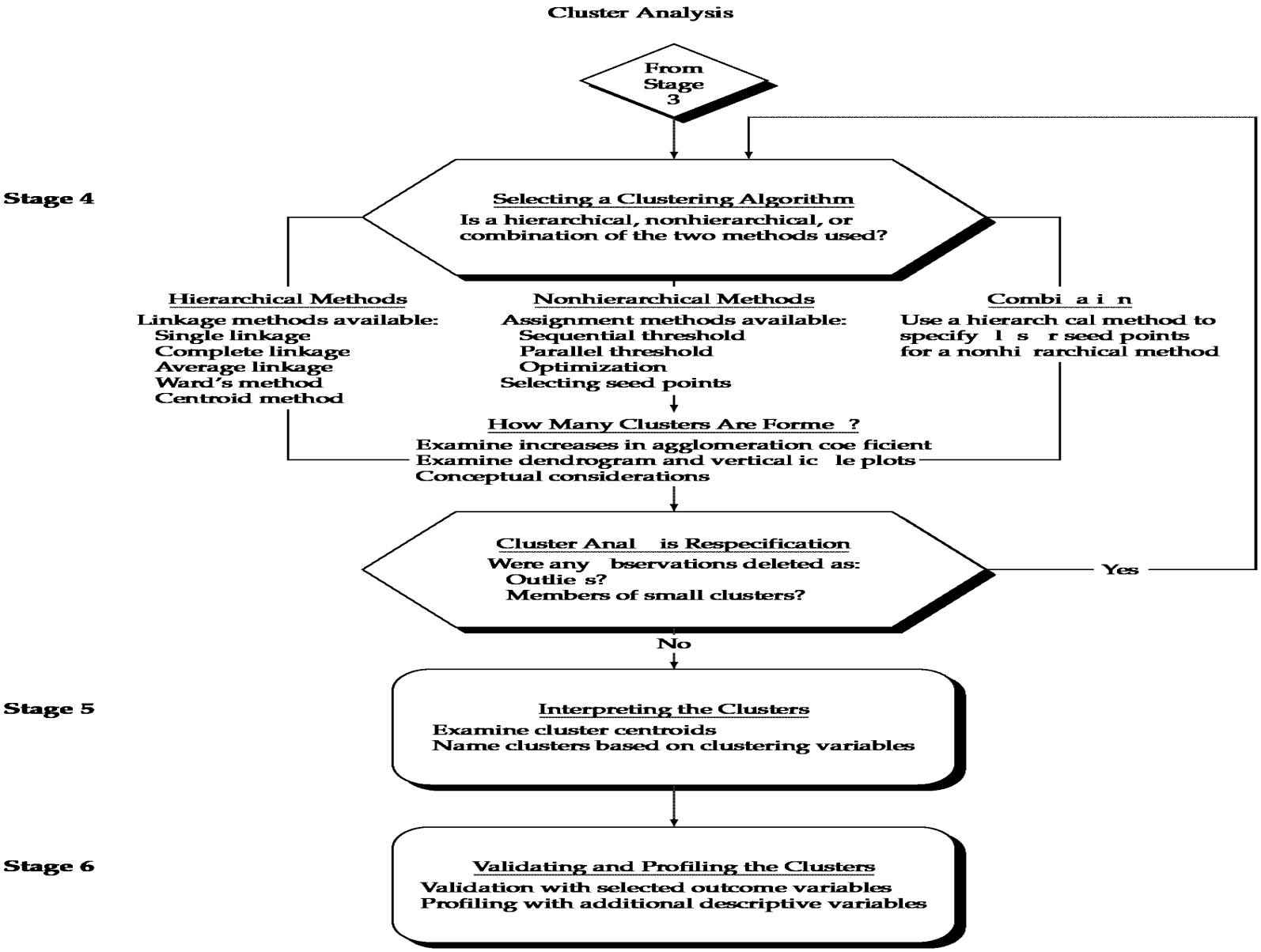


FIGURE 6 *Stages 4–6 of the Cluster Analysis Decision Diagram*

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - With the clustering variables selected and the similarity matrix calculated, the partitioning process begins.
  - • Select the partitioning procedure used for forming clusters.
  - • Make the decision on the number of clusters to be formed.
  - Both decisions have substantial implications not only on the results that will be obtained but also on the interpretation that can be derived from the results .
  - First, we examine the available partitioning procedures and then discuss the options available for deciding on a cluster solution by defining the number of clusters and membership for each observation.
  - Partitioning procedures work on a simple principle. They seek to maximize the distance between groups while minimizing the differences of in-group members.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**

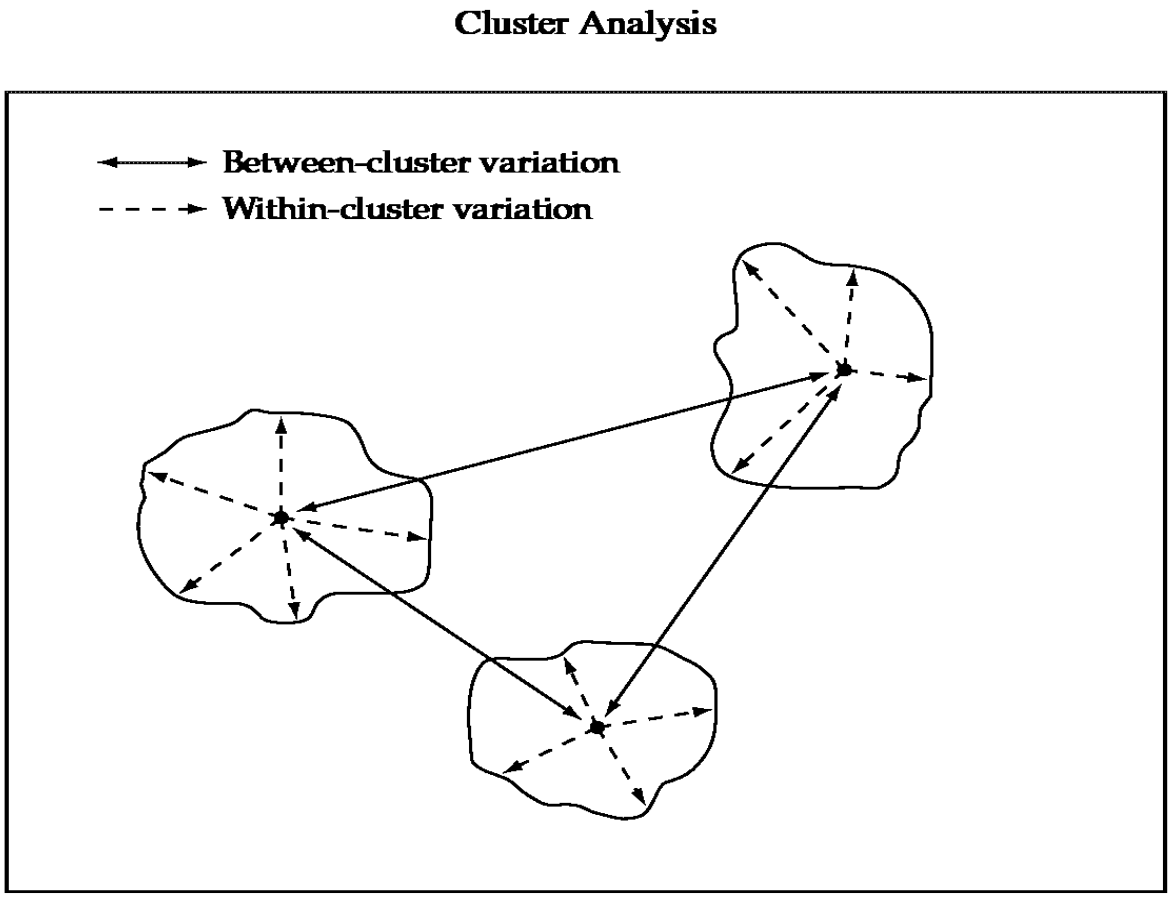


FIGURE 7 Cluster Diagram Showing Be w en- and Within-Cluster Variation

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - Hierarchical procedures involve a series of n – 1 clustering decisions (where n equals the number of observations) that combine observations into a hierarchy or a treelike structure.
  - The two basic types of hierarchical clustering procedures are agglomerative and divisive.
  - In the **agglomerative methods**, each object or observation starts out as its own cluster and is successively joined, the two most similar clusters at a time until only a single cluster remains.
  - In **divisive methods** all observations start in a single cluster and are successively divided (first into two clusters, then three, and so forth) until each is a single-member cluster.
  - In Figure 8, agglomerative methods move from left to right, and divisive methods move from right to left.
  - we focus here on the agglomerative methods.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**

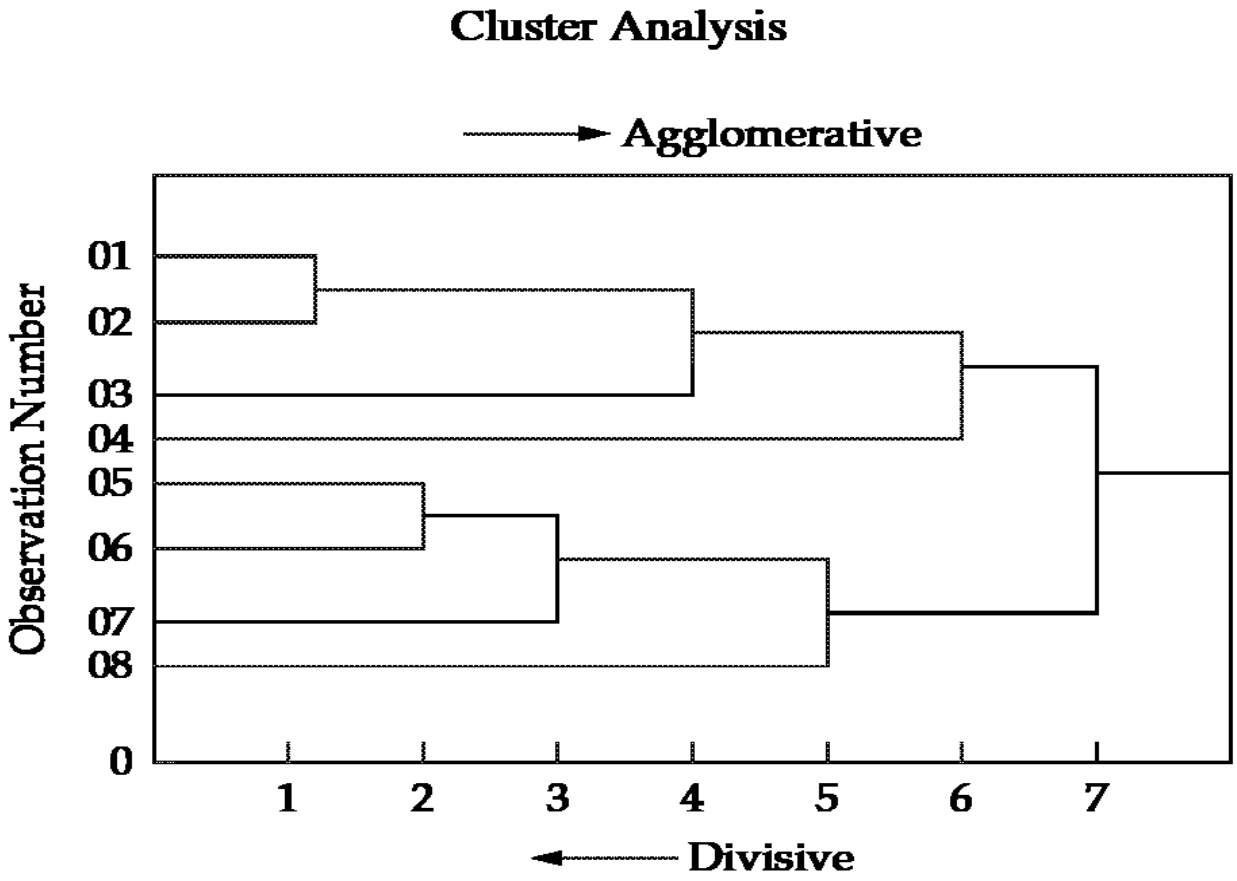


FIGURE 8 Dendrogram Illustrating Hierarchical Clustering

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - To understand how a hierarchical procedure works, we will examine the most common form—the agglomerative method—which follows a simple, repetitive process:
  - 1. Start with all observations as their own cluster (i.e., each observation forms a single-member cluster), so that the number of clusters equals the number of observations.
  - 2. Using the similarity measure, combine the two most similar clusters into a new cluster (now containing two observations), thus reducing the number of clusters by one.
  - 3. Repeat the clustering process again, using the similarity measure to combine the two most similar clusters into a new cluster.
  - 4. Continue this process, at each step combining the two most similar clusters into a new cluster. Repeat the process a total of n – 1 times until all observations are contained in a single cluster.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - Assume that we had 100 observations.
  - We would initially start with 100 separate clusters, each containing a single observation.
  - At the first step, the two most similar clusters would be combined, leaving us with 99 clusters.
  - At the next step, we combine the next two most similar clusters, so that we then have 98 clusters.
  - This process continues until the last step where the final two remaining clusters are combined into a single cluster.
  - An important characteristic of hierarchical procedures is that the results at an earlier stage are always nested within the results at a later stage, creating a similarity to a tree.
  - For example, an agglomerative 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 8; the representation is referred to as a **dendrogram or tree graph,** which can be useful, but becomes unwieldy with large applications.
  - The dendrogram is widely available in most clustering software.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **Clustering Algorithms.**
  - The clustering algorithm in a hierarchical procedure defines how similarity is defined between multiple-member clusters in the clustering process.
  - So how do we measure similarity between clusters when one or both clusters have multiple members?
  - Do we select one member to act as a typical member and measure similarity between these members of each cluster, or do we create some composite member to represent the cluster, or even combine the similarities between all members of each cluster?
  - We could employ any of these approaches, or even devise other ways to measure similarity between multiple-member clusters.
  - Among numerous approaches, the five most popular agglomerative algorithms are (1) Single-linkage
  - (2) Complete-linkage
  - (3) Average linkage
  - (4) Centroid method and
  - (5) Ward's method.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **Single-Linkage**
  - The single-linkage method (also called the nearest-neighbor method) defines the similarity between clusters as the shortest distance from any object in one cluster to any object in the other.
  - This rule was applied in the example at the beginning of this chapter and enables us to use the original distance matrix between observations without calculating new distance measures.
  - Just find all the distances between observations in the two clusters and select the smallest as the measure of cluster similarity.
  - This method is probably the most versatile agglomerative algorithm, because it can define a wide range of clustering patterns (e.g., it can represent clusters that are concentric circles, like rings of a bull's-eye).

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - An example of this arrangement is shown in Figure 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 figure shows the principal disadvantage of the single-linkage algorithm.

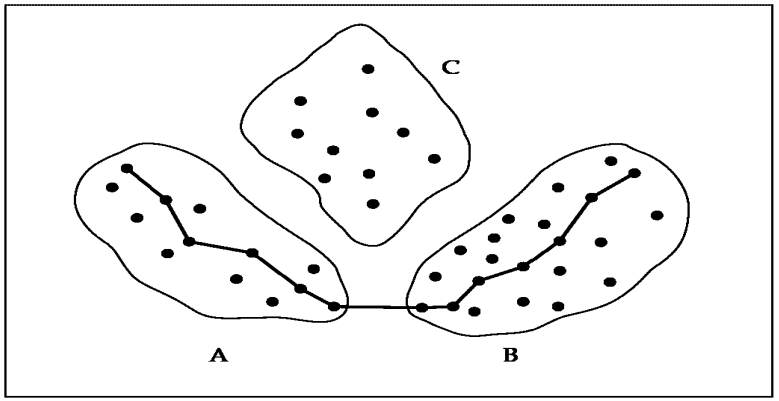


FIGURE 9 Example of Single Linkage Joining Dissimilar Clusters A and B

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **• Complete-Linkage**
  - The complete-linkage method (also known as the farthest-neighbor or diameter method) is comparable to the single-linkage algorithm, except that cluster similarity is based on maximum distance between observations in each cluster.
  - Similarity between clusters is the smallest (minimum diameter) sphere that can enclose all observations in both clusters.
  - This method is called complete-linkage because all objects in a cluster are linked to each other at some maximum distance.
  - Thus, within-group similarity equals group diameter. This technique eliminates the chaining problem identified with single-linkage and has been found to generate the most compact clustering solutions.
  - Even though it represents only one aspect of the data (i.e., the farthest distance between members), many researchers find it the most appropriate for a wide range of clustering applications.
  - Figure 10 compares the shortest (single-linkage) and longest (complete-linkage) distances in representing similarity between clusters.
  - Both measures reflect only one aspect of the data.
  - The use of the single-linkage reflects only a closest single pair of objects, and the complete-linkage also reflects a single pair, this time the two most extreme.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**

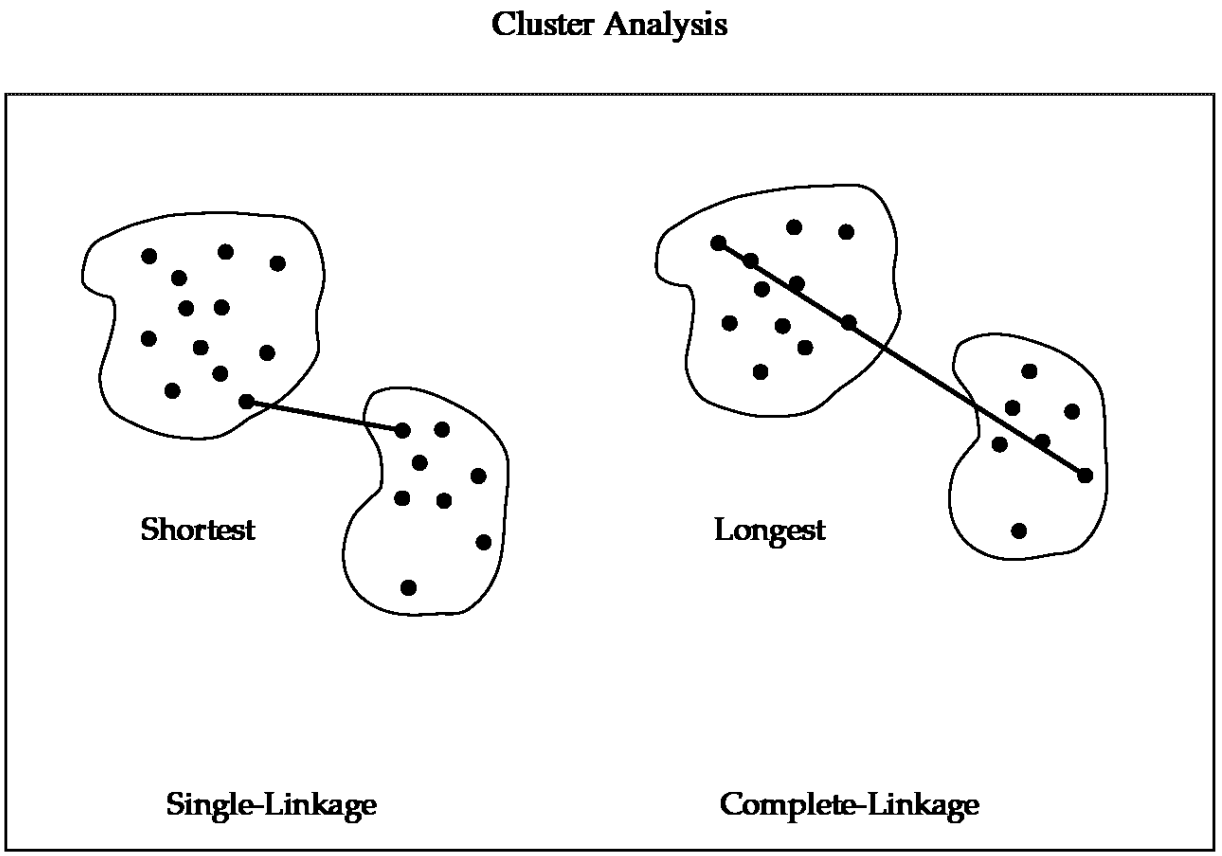


FIGURE 10 Comparison of Distance Measures for Single-Linkage and Complete-Linkage

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **• Average Linkage**
  - The average linkage procedure differs from the single-linkage or complete-linkage procedures in that the similarity of any two clusters is the average similarity of all individuals in one cluster with all individuals in another.
  - This algorithm does not depend on extreme values (closest or farthest pairs) as do single-linkage or complete-linkage.
  - Instead, similarity is based on all members of the clusters rather than on a single pair of extreme members and are thus less affected by outliers.
  - Average linkage approaches, as a type of compromise between single- and complete-linkage methods, tend to generate clusters with small within-cluster variation.
  - They also tend toward the production of clusters with approximately equal within-group variance.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **• Centroid Method**
  - In the centroid method the similarity between two clusters is the distance between the cluster 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, a cluster centroid changes every time a new individual or group of individuals is added to an existing cluster.
  - These methods are the most popular in the physical and life sciences (e.g., biology) but 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, like the average linkage method, is that it is less affected by outliers than are other hierarchical methods.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **• Ward's Method**
  - The Ward's method differs from the previous methods in that the similarity between two clusters is not a single measure of similarity, but rather the sum of squares within the clusters summed over all variables.
  - It is quite similar to the simple heterogeneity measure used in the example at the beginning of the chapter to assist in determining the number of clusters.
  - In the Ward's procedure, the selection of which two clusters to combine is based on which combination of clusters minimizes the within-cluster sum of squares across the complete set of disjoint or separate clusters.
  - At each step, the two clusters combined are those that minimize the increase in the total sum of squares across all variables in all clusters.
  - This procedure tends to combine clusters with a small number of observations, because the sum of squares is directly related to the number of observations involved.
  - The use of a sum of squares measure makes this method easily distorted by outliers.
  - Moreover, the Ward's method also tends to produce clusters with approximately the same number of observations.
  - However, the use of this method also makes it more difficult to identify clusters representing small proportions of the sample.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HIERARCHICAL CLUSTER PROCEDURES**
  - **• Overview**
  - Hierarchical clustering procedures are a combination of a repetitive clustering process combined with a clustering algorithm to define the similarity between clusters with multiple members.
  - The process of creating clusters generates a treelike diagram that represents the combinations/divisions of clusters to form the complete range of cluster solutions.
  - Hierarchical procedures generate a complete set of cluster solutions, ranging from all single-member clusters to the one-cluster solution where all observations are in a single cluster.
  - In doing so, the hierarchical procedure provides an excellent framework with which to compare any set of cluster solutions and help in judging how many clusters should be retained.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **NON-HIERARCHICAL 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 is specified.
  - For example, a 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.
  - The nonhierarchical cluster software programs usually proceed through two steps:
  - **1. Specify cluster seeds:**
  - The first task is to identify starting points, known as cluster seeds, for each cluster. A cluster seed may be pre-specified by the researcher or observations may be selected, usually in a random process.
  - **2. Assignment:**
  - With the cluster seeds defined, the next step is to assign each observation to one of the cluster seeds based on similarity.
  - Many approaches are available for making this assignment (see later discussion in this section), but the basic objective is to assign each observation to the most similar cluster seed.
  - In some approaches, observations may be reassigned to clusters that are more similar than their original cluster assignment.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **NON-HIERARCHICAL CLUSTERING PROCEDURES**
  - We discuss several different approaches for selecting cluster seeds and assigning objects in the next sections.
  - **Selecting Seed Points**
  - Even though the nonhierarchical clustering algorithms discussed in the next section differ in the manner in which they assign observations to the seed points, they all face the same problem:
  - How do we select the cluster seeds?
  - The different approaches can be classified into two basic categories:
  - **1. Researcher specified.**
  - In this approach, the researcher provides the seed points based on external data.
  - The two most common sources of the seed points are prior research or data from another multivariate analysis.
  - Many times the researcher has knowledge of the cluster profiles being researched.
  - For example, prior research may have defined segment profiles and the task of the cluster analysis is to assign individuals to the most appropriate segment cluster.
  - It is also possible that other multivariate techniques may be used to generate the seed points.
  - One common example is the use of a hierarchical clustering algorithm to establish the number of clusters and then generate seed points from these results.
  - The common element is that the researcher, while knowing the number of clusters to be formed, also has information about the basic character of these clusters.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **NON-HIERARCHICAL CLUSTERING PROCEDURES**
  - **2. Sample generated.**
  - The second approach is to generate the cluster seeds from the observations of the sample, either in some systematic fashion or simply through random selection.
  - For example, in the FASTCLUS program in SAS, 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 are selected, the program assigns each observation to the cluster with the nearest seed.
  - The K-means program in SPSS can select the necessary seed points randomly from among the observations. In any of these approaches, the researcher relies on the selection process to choose seed points that reflect natural clusters as starting points for the clustering algorithms.
  - A possible limitation is that replication of the results is difficult if the observations are reordered.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **NON-HIERARCHICAL CLUSTERING PROCEDURES**
  - In either approach, the researcher must be aware of the impact of the cluster seed selection process on the final results.
  - All of the clustering algorithms, even those of an optimizing nature (see the following discussion), will generate different cluster solutions depending on the initial cluster seeds.
  - The differences among cluster solutions will hopefully be minimal using different seed points, but they underscore the importance of cluster seed selection and its impact on the final cluster solution.
  - **Nonhierarchical Clustering Algorithms.**
  - Several clustering algorithms have been proposed. The most frequently cited are sequential, parallel and optimization.
  - The sequential threshold method starts by selecting one cluster seed and includes all objects within a pre-specified distance.
  - A second cluster seed is then selected, and all objects within the pre-specified distance of that seed are included.
  - A third seed is then selected, and the process continues as before.
  - The primary disadvantage of this approach is that when an observation is assigned to a cluster, it cannot be reassigned to another cluster, even if that cluster seed is more similar.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **NON-HIERARCHICAL CLUSTERING PROCEDURES**
  - The parallel threshold method considers all cluster seeds simultaneously and assigns observations within the threshold distance to the nearest seed.
  - The third method, referred to as the optimizing procedure, is similar to the other two nonhierarchical procedures except that it allows for reassignment of observations to a seed other than the one with which it was originally associated.
  - All of these belong to a group of clustering algorithms known as **K-means**.
  - **K-means algorithms** work by portioning the data into a user-specified number of clusters and then iteratively reassigning observations to clusters until some numerical criterion is met.
  - The criterion specifies a goal related to minimizing the distance of observations from one another in a cluster and maximizing the distance between clusters.
  - K-means is so commonly used that the term is used by some to refer to nonhierarchical cluster analysis in general.
  - For example, in SPSS, the nonhierarchical clustering routine is referred to as K-means.
  - An optimizing procedure allows for reassignment of observations based on the goal of creating the most distinct clusters.
  - If, in the course of assigning observations, an observation becomes closer to another cluster that is not the cluster to which it is currently assigned, then an optimizing procedure switches the observation to the more similar (closer) cluster.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - We can examine the strengths and weaknesses of each method to determine which is most appropriate for a unique research setting.
  - **Pros and Cons of Hierarchical Methods.**
  - Hierarchical clustering techniques have long been the more popular clustering method, with Ward's method and average linkage probably being the best available.
  - Besides the fact that hierarchical procedures were the first clustering methods developed, they do offer several advantages that lead to their widespread usage:
  - **1. Simplicity:**
  - Hierarchical techniques, with their development of the treelike structures depicting the clustering process, do afford the researcher with a simple, yet comprehensive, portrayal of the entire range of clustering solutions.
  - In doing so, the researcher can evaluate any of the possible clustering solutions from one analysis.
  - **2. Measures of similarity:**
  - The widespread use of the hierarchical methods led to an extensive development of similarity measures for almost any type of clustering variables.
  - Hierarchical techniques can be applied to almost any type of research question.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - **3. Speed:**
  - Hierarchical methods have the advantage of generating an entire set of clustering solutions (from all separate clusters to one cluster) in an expedient manner.
  - This ability enables the researcher to examine a wide range of alternative clustering solutions, varying measures of similarities and linkage methods, in an efficient manner.
  - Even though hierarchical techniques have been widely used, they do have several distinct **disadvantages** that affect any of their cluster solutions:
  - 1. Hierarchical methods can be misleading because undesirable early combinations may persist throughout the analysis and lead to artificial results. Of specific concern is the substantial impact of outliers on hierarchical methods, particularly with the complete-linkage method.
  - 2. To reduce the impact of outliers, the researcher may wish to cluster analyze the data several times, each time deleting problem observations or outliers.
  - The deletion of cases, however, even those not found to be outliers, can many times distort the solution.
  - Thus, the researcher must employ extreme care in the deletion of observations for any reason.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - 3. Although computations of the clustering process are relatively fast, hierarchical methods are not amenable to analyzing large samples or even large numbers of variables.
  - As sample size increases, the data storage requirements increase dramatically.
  - For example, a sample of 400 cases requires storage of approximately 80,000 similarities, which increases to almost 125,000 for a sample of 500.
  - Even with the computing power of today's personal computers, these data requirements can limit the application in many instances.
  - The researcher may take a random sample of the original observations to reduce sample size but must now question the representativeness of the sample taken from the original sample.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - **Emergence of Nonhierarchical Methods.**
  - Nonhierarchical methods have gained increased acceptability and usage, but any application 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 offer several advantages over hierarchical techniques.
  - 1. The results are less susceptible to outliers in the data, the distance measure used, and the inclusion of irrelevant or inappropriate variables.
  - 2. Nonhierarchical methods can analyze extremely large data sets because they do not require the calculation of similarity matrices among all observations, but instead just the similarity of each observation to the cluster centroids.
  - Even the optimizing algorithms that allow for reassignment of observations between clusters can be readily applied to all sizes of data sets.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - **Emergence of Nonhierarchical Methods.**
  - Although nonhierarchical methods do have several distinct advantages, several shortcomings can markedly affect their use in many types of applications.
  - 1. The benefits of any nonhierarchical method are realized only with the use of nonrandom (i.e., specified) seed points. Thus, the use of nonhierarchical techniques with random seed points is generally considered inferior to hierarchical techniques.
  - 2. Even a nonrandom starting solution does not guarantee an optimal clustering of observations.
  - In fact, in many instances the researcher will get a different final solution for each set of specified seed points.
  - Only by analysis and validation can the researcher select what is considered the best representation of structure, realizing that many alternatives may be as acceptable.
  - 3. Nonhierarchical methods are also not as efficient when examining a large number of potential cluster solutions.
  - Each cluster solution is a separate analysis, in contrast to the hierarchical techniques that generate all possible cluster solutions in a single analysis.
  - Thus, nonhierarchical techniques are not as well suited to exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **SHOULD HIERARCHICAL OR NONHIERARCHICAL METHODS BE USED?**
  - **A Combination of Both Methods.**
  - Many researchers recommend a combination approach using both methods. In this way, the advantages of one approach can compensate for the weaknesses or the other. This is accomplished in two steps:
  - 1. First, a hierarchical technique is used to generate a complete set of cluster solutions, establish the applicable cluster solutions, and establish the appropriate number of clusters.
  - 2. After outliers are eliminated, the remaining observations can then be clustered by a nonhierarchical method. In this way, the advantages of the hierarchical methods are complemented by the ability of the nonhierarchical methods to refine the results by allowing the switching of cluster membership.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **HOW MANY CLUSTERS SHOULD BE FORMED?**
  - Stopping rule that suggests two or more cluster solutions which can be compared before making the final decision.
  - Unfortunately, no standard objective selection procedure exists. Because no internal statistical criterion is used for inference, such as the statistical significance tests of other multivariate methods, researchers have developed many criteria for approaching the problem.
  - The principal issues facing any of these stopping rules include the following:
  - • These ad hoc procedures must be computed by the researcher and often involve fairly complex.
  - • Many of these criteria are specific to a particular software program and are not easily calculated if not provided by the program.
  - • A natural increase in heterogeneity comes from the reduction in number of clusters. Thus, the researcher must look at the trend in the values of these stopping rules across cluster solutions to identify marked increases. If not, in most instances the two-cluster solution would always be chosen because the value of any stopping rule is normally highest when going from two to one cluster.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **Measures of Heterogeneity Change.**
  - One class of stopping rules examines some measure of heterogeneity between clusters at each successive step, with the cluster solution defined when the heterogeneity measure exceeds a specified value or when the successive values between steps makes a sudden jump.
  - Heterogeneity refers to how different the observations in a cluster are from each other (i.e., heterogeneity refers to a lack of similarity among group members).
  - A simple example was used 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 increase in heterogeneity.
  - This type of stopping rule has been shown to provide fairly accurate decisions in empirical studies, but it is not uncommon for a number of cluster solutions to be identified by these large increases in heterogeneity.
  - It is then the researcher's task to select a final cluster solution from these selected cluster solutions. Various stopping rules follow this general approach.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **Percentage Changes in Heterogeneity**
  - Probably the simplest and most widespread rule is a simple percentage change in some measure of heterogeneity.
  - A typical example is using the agglomeration coefficient in SPSS, which measures heterogeneity as the distance at which clusters are formed (if a distance measure of similarity is used) or the within-cluster sum of squares if the Ward's method is used.
  - With this measure, the percentage increase in the agglomeration coefficient can be calculated for each cluster solution.
  - Then the researcher selects cluster solutions as a potential final solution when the percentage increase is markedly larger than occurring at other steps.
  - **Measures of Variance Change**
  - The root mean square standard deviation (RMSSTD) is the square root of the variance of the new cluster formed by joining the two clusters.
  - The variance for the newly formed cluster is calculated as the variance across all clustering variables.
  - Large increases in the RMSSTD suggest the joining of two quite dissimilar clusters, indicating the previous cluster solution (in which the two clusters were separate) was a candidate for selection as the final cluster solution.

Cluster Analysis Decision Process

- **Stage 4: Deriving Clusters and Assessing Overall Fit**
  - **Statistical Measures of Heterogeneity Change**
  - A series of test statistics attempts to portray the degree of heterogeneity for each new cluster solution (i.e., joining of two clusters).
  - One of the most widely used is a pseudo F statistic, which compares the goodness-of-fit of k clusters to k – 1 clusters.
  - Highly significant values indicate that the k – 1 cluster solution is more appropriate than the k cluster solution.
  - The researcher should not consider any significant value, but instead look to those values markedly more significant than for other solutions.
  - **Direct Measures of Heterogeneity.**
  - A second general class of stopping rules attempts to directly measure heterogeneity of each cluster solution.
  - The most common measure in this class is the cubic clustering criterion (CCC) contained in SAS, a measure of the deviation of the clusters from an expected distribution of points formed by a multivariate uniform distribution.
  - Here the researcher selects the cluster solution with the largest value of CCC (i.e., the cluster solution where CCC peaks).
  - Despite its inclusion in SAS and its advantage of selecting a single-cluster solution, it has been shown to many times generate too many clusters as the final solution and is based on the assumption that the variables are uncorrelated.
  - However, it is a widely used measure and is generally as efficient as any other stopping rule.

Cluster Analysis Decision Process

- **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.
  - Let us assume that an attitude scale was developed that consisted of statements regarding consumption of soft drinks, such as “diet soft drinks taste harsher,” “regular soft drinks have a fuller taste,” “diet drinks are healthier,” and so forth.
  - Further, let us assume that demographic and softdrink consumption data were also collected.
  - When starting the interpretation process, one measure frequently used is the cluster’s centroid.
  - If the clustering procedure was performed on the raw data, it would be a logical description.
  - If the data were standardized or if the cluster analysis was performed using factor analysis (component factors), the researcher would have to go back to the raw scores for the original variables and develop profile diagrams using these data.

Cluster Analysis Decision Process

- **Stage 5: Interpretation of the Clusters**
  - The profiling and interpretation of the clusters, however, achieve more than just description and are essential elements in selecting between cluster solutions when the stopping rules indicate more than one appropriate cluster solution.
  - • 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.
  - • The cluster profiles also 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 be expanded until such differences arise.
  - In assessing either correspondence or practical significance, the researcher compares the derived clusters to a preconceived typology.
  - This more subjective judgment by the researcher combines with the empirical judgment of the stopping rules to determine the final cluster solution to represent the data structure of the sample.

Cluster Analysis Decision Process

- **Stage 6: Validation and Profiling of the Clusters**
  - **VALIDATING THE CLUSTER SOLUTION**
  - Validation includes attempts by the researcher to assure that the cluster solution is representative of the general population, and thus is generalizable to other objects and is stable over time.
  - **Cross-Validation.**
  - 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 objects (particularly consumers) for multiple cluster analyses.
  - In these instances, a common approach is to split the sample into two groups.
  - Each cluster is analyzed separately, and the results are then compared.
  - Cross-tabulation also can be used, because the members of any specific cluster in one solution should stay together in a cluster in another solution.
  - Therefore, the cross-tabulation should display obvious patterns of matching cluster membership.
  - Other approaches include
  - (1) a modified form of split sampling whereby cluster centers obtained from one cluster solution are employed to define clusters from the other observations and the results are compared, and
  - (2) a direct form of cross-validation.

Cluster Analysis Decision Process

- **Stage 6: Validation and Profiling of the Clusters**
  - **VALIDATING THE CLUSTER SOLUTION**
  - For any of these methods, stability of the cluster results can be assessed by the number of cases assigned to the same cluster across cluster solutions.
  - Generally, a very stable solution would be produced with less than 10 percent of observations being assigned to a different cluster.
  - A stable solution would result with between 10 and 20 percent assigned to a different group, and a somewhat stable solution when between 20 and 25 percent of the observations are to a different cluster than the initial one.
  - **Establishing Criterion Validity.**
  - The researcher may also attempt to establish some form of criterion or predictive validity.
  - To do so, the researcher selects variable(s) 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 that are favorable to diet soft drinks and those that are not.
  - The variable(s) used to assess predictive validity should have strong theoretical or practical support because they become the benchmark for selecting among the cluster solutions.

Cluster Analysis Decision Process

- **Stage 6: Validation and Profiling of the Clusters**
  - **PROFILING THE CLUSTER SOLUTION**
  - The profiling stage involves describing the characteristics of each cluster to explain how it may differ on relevant dimensions.
  - This process typically involves the use of discriminant analysis. The procedure begins after the clusters are identified.
  - The researcher utilizes data not previously included in the cluster procedure to profile the characteristics of each cluster.
  - These data typically are demographic characteristics, psychographic profiles, consumption patterns, and so forth.
  - Although no theoretical rationale may exist for their difference across the clusters, such as required for predictive validity assessment, they should at least have practical importance.
  - Using discriminant analysis, the researcher compares 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.

Cluster Analysis Decision Process

- **Stage 6: Validation and Profiling of the Clusters**
  - **PROFILING THE CLUSTER SOLUTION**
  - From this analysis, assuming statistical significance, the researcher could conclude, for example, that the "health- and calorie-conscious" cluster from our previous diet soft drink 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 rather on the characteristics of the clusters after they are identified.
  - Moreover, the emphasis is on the characteristics that differ significantly across the clusters and those that could predict membership in a particular cluster.
  - Profiling often is an important practical step in clustering procedures, because identifying characteristics like demographics enables segments to be identified or located with easily obtained information.