Chapter 10
Cluster Analysis: For Segmenting the Population

Learning Objectives

After completing this chapter, you should be able to do the following:

- Understand the concept of cluster analysis.
- Know the different terminologies used in cluster analysis.
- Learn to compute different distances used in the analysis.
- Understand different techniques of clustering.
- Describe the assumptions used in the analysis.
- Explain the situations where cluster analysis can be used.
- Learn the procedure of using cluster analysis.
- Know the use of hierarchical cluster analysis and *K*-means cluster analysis.
- Describe the situation under which two-step cluster should be used.
- Understand various outputs of cluster analysis.
- Know the procedure of using cluster analysis with SPSS.
- Understand different commands and its outcomes used in SPSS for cluster analysis.
- Learn to interpret the outputs of cluster analysis generated by the SPSS.

Introduction

Market analysts are always in search of strategies responsible for buying behavior. The whole lot of customers can be grouped on the basis of their buying behavior patterns. This segmentation of customers helps analysts in developing marketing strategy for different products in different segments of customers. These segments are developed on the basis of buying behavior of the customers in such a way so that the individuals in the segments are more alike but the individuals in different segments differ to a great extent in their characteristics. The concept of segmenting may be used to club different television serials into homogeneous categories on the basis of their characteristics. An archaeological surveyor's may like to cluster different idol excavated from archaeological digs into the civilizations from which

J.P. Verma, *Data Analysis in Management with SPSS Software*,
DOI 10.1007/978-81-322-0786-3_10, © Springer India 2013

they originated. These idols may be clustered on the basis of their physical and chemical parameters to identify their age and civilization to which they belong. Doctors may diagnose a patient for viral infection and determine whether distinct subgroups can be identified on the basis of a clinical checklist and pathological tests. Thus, in different fields several situations may arise where it is required to segment the subjects on the basis of their behaviour pattern so that an appropriate strategy may be formed for these segments separately. Segmenting may also be done for the objects based on their similarity of features and characteristics. Such segmenting of objects may be useful for making a policy decision. For instance, all the cars can be classified into small, medium and large segments depending upon their features like engine power, price, seating capacity, luggage capacity, and fuel consumption. Different policy may be adopted to promote these segments of vehicle by the authorities.

The problem of segmentation shall be discussed in this chapter by means of cluster analysis. The more emphasis has been given on understanding various concepts of this analysis and the procedure used in it. Further, solved example has been discussed by means of using SPSS for easy understanding of readers. The reader should note as to how different outputs generated in this analysis by the SPSS have been interpreted.

What Is Cluster Analysis?

Cluster analysis is a multivariate statistical technique for grouping cases of data based on the similarity of responses to several variables/subjects. The purpose of cluster analysis is to place subjects/objects into groups, or clusters, suggested by the data, such that objects in a given cluster are homogenous in some sense, and objects in different clusters are dissimilar to a great extent. In cluster analysis, the groups are not predefined but are rather suggested on the basis of the data. The cluster analysis can also be used to summarize data rather than to find observed clusters. This process is sometimes called dissection.

Terminologies Used in Cluster Analysis

Distance Measure

In cluster analysis, cases/objects are clustered on the basis of dissimilarities (similarities) or distances between cases/objects. These distances (similarities) can be based on a single or multiple parameters where each parameter represents a rule or condition for grouping cases/objects. For example, if we were to cluster the songs, we may take into account the song length, singer, subjective ratings of the

Table 10.1 Employees' profile

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Employee 3	1.2	1.9	−0.9
Employee 4	1.5	−0.4	1.3

listeners, etc. The simplest way of computing distances between cases in a multidimensional space is to compute Euclidean distances. There are many methods available for computing distances and it is up to the researcher to identify an appropriate method according to the nature of the problem. Although plenty of methods are available for computing distances between the cases, we are discussing herewith the five most frequently used methods. These methods for computing the distances shall be discussed later in this chapter by using some data.

Consider the data in Table 10.1 where age, income, and qualification are the three different parameters on which employees need to be grouped into different clusters. We will see the computation of distances between the two employees using different distance method.

Squared Euclidean Distance

A Euclidean distance is a geometric distance between two cases or objects. This is the most natural way of computing a distance between two samples. It computes the difference between two samples directly on the basis of the changes in magnitude in the sample levels. Euclidean distance is usually used in a situation where data sets are suitably normalized. It is computed by taking the square root of the sum of the squared difference on each of the variable measurements between the two cases. The formula for its computation is given by

$$\mathrm{de}_{ij} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2} \tag{10.1}$$

where

X_{ik} is the measurement of ith cases on kth variable
X_{jk} is the measurement of jth cases on kth variable
n is number of variables

Let us compute the Euclidean distance between first and second employee by using their profile as shown in Table 10.1.

Table 10.2 Computation of Euclidean space between employees 1 and 2

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Difference	0.2	0.3	0.5
Squared difference	0.04	0.09	0.25

Table 10.3 Computation of Manhattan distance between employees 1 and 2

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Absolute difference	0.2	0.3	0.5

The squared Euclidean distance between employee 1 and employee 2 can be obtained by using the formula 10.1. The computation has been shown in the Table 10.2.

$$\text{Thus, Squared Euclidean space between first and second employee} = 0.04 + 0.09 + 0.25 = 0.38$$

Since Euclidean distance is the square root of the squared Euclidean distance,

Euclidean distance between first and second employee $= \text{de}_{12} = \sqrt{0.38} = 0.62$.

In computing the Euclidean distance, each difference is squared to find the absolute difference on each of the variables measured on both the employees. After adding all of the squared differences, we take the square root. We do it because by squaring the differences, the units of measurements are changed, and so by taking the square root, we get back the original unit of measurement.

If Euclidean distances are smaller, the cases are more similar. However, this measure depends on the units of measurement for the variables. If variables are measured on different scales, variables with large values will contribute more to the distance measure than the variables with small values. It is therefore important to standardize scores before proceeding with the analysis if variables are measured on different scales. In SPSS, you can standardize variables in different ways.

Manhattan Distance

The Manhattan distance between the two cases is computed by summing the absolute distances along each variable. The Manhattan distance is also known as city-block distance and is appropriate when the data set is discrete. By using the data of Table 10.1 the Manhattan distance between first and second employee has been computed in the Table 10.3.

Thus, the Manhattan distance between first and second employees $= \text{dm} = 0.2 + 0.3 + 0.5 = 1.00$.

Chebyshev Distance

The Chebyshev distance between the two cases are obtained by finding the maximum absolute difference in values for any variable. This distance is computed if we want to define two cases as "different" if they differ on any one of the dimensions. The Chebyshev distance is computed as

$$\text{Chebyshev distance } (x, y) = dc = \text{Max } |x_i - y_i| \tag{10.2}$$

In Table 10.1, the Chebyshev distance between the first and fourth employees would be 2.8 as this is the maximum absolute difference of these two employees on income variable.

Mahalanobis (or Correlation) Distance

The Mahalanobis distance is based on the Pearson correlation coefficient which is computed between the observations of two cases or subjects. This correlation coefficient is used to cluster the cases. This is an important measure as it is a scale invariant. In other words, it is not affected by the change in units of the observations. Thus, the Mahalanobis distance (dm) between first and second employees can be obtained by computing the correlation coefficient between the observations 2.5, 2.4, 2.4 and 2.3, 2.1, 1.9.

Pearson Correlation Distance

The Pearson distance (dp) is also based on the Pearson correlation coefficient between the observations of the two cases. This distance is computed as $dp = 1-r$ and lies between 0 and 2. Since the maximum and minimum values of r can be +1 and -1, respectively, the range of the Pearson distance (dp) can be from 0 to 2. The zero value of dp indicates that the cases are alike, and 2 indicates that the cases are entirely distinct.

Clustering Procedure

In cluster analysis, each case/object is considered to be a single cluster. The distances between these objects are computed by the chosen distance measure. On the basis of these distances computed in the proximity matrix, several objects are linked together. After having done so, how do we determine the distances between these new clusters? In other words, we need to have a linkage or amalgamation criteria to determine when two clusters are sufficiently similar to be linked together. There are various protocols: for example, we may link two clusters

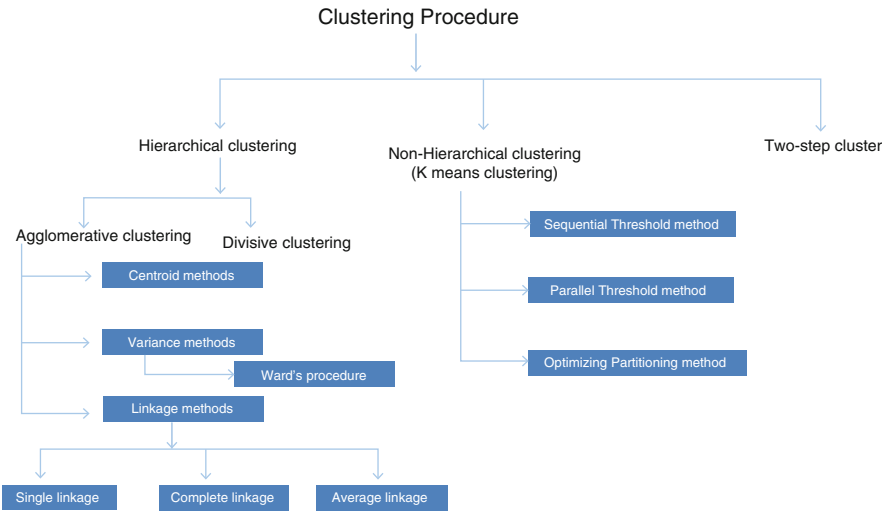


Fig. 10.1 Different clustering procedures

together on the basis of the smallest distance between the two objects, one from each of the two different clusters. Similarly the two clusters may be linked together on the basis of the maximum distance between the two objects, one from each cluster. There are different ways the objects can be clustered together. The entire clustering procedures can be broadly classified in three different categories, that is, hierarchical clustering, nonhierarchical clustering, and two-step clustering. These procedures shall be discussed in detail under various headings in this section. The details of various classification procedures have been shown graphically in Fig. 10.1.

Hierarchical Clustering

In hierarchical clustering, objects are organized into a hierarchical structure. It creates a hierarchy of clusters which may be represented in a treelike structure known as dendrogram. Objects are grouped into a tree of clusters by using the distance (similarity) matrix as clustering criteria. In this tree structure, the root consists of a single cluster containing all observations, whereas the leaves refer to the individual observations. *Hierarchical clustering* is the best for small data sets because in this procedure a proximity matrix of the distance/similarity is computed for each pair of cases in the data set.

Hierarchical clustering can be either agglomerative or divisive. In agglomerative clustering, one starts at the individual objects and successively merges clusters together. On the other hand, in the divisive clustering, one starts with all the objects as one cluster and recursively splits the clusters. We shall now discuss various types of clustering protocols of these two types of hierarchical clustering in detail.

Fig. 10.2 Raw data showing the distances between the objects

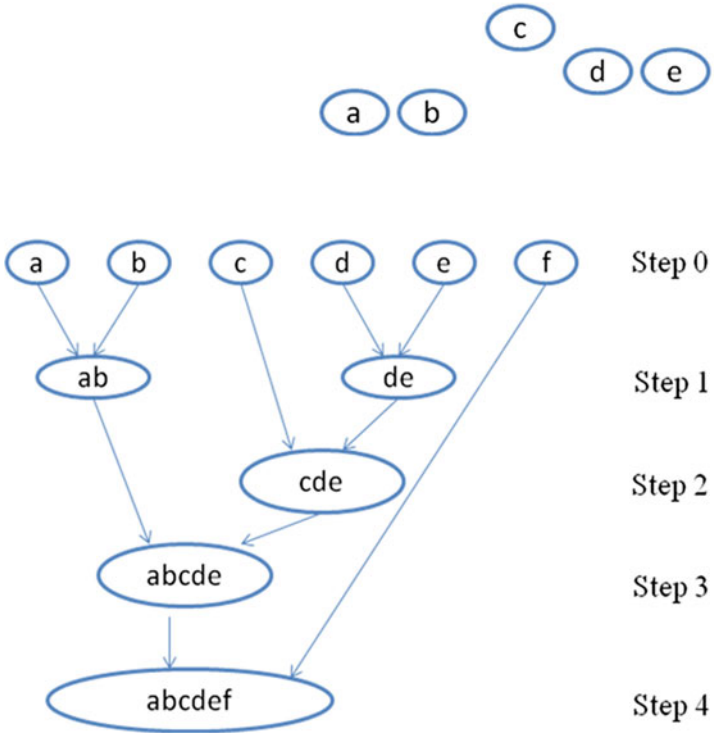


Fig. 10.3 Formation of clusters at different stages in agglomerative clustering

Agglomerative Clustering

In agglomerative clustering, all the individual objects/cases are considered as a separate cluster. These objects (atomic clusters) are successively merged into bigger and bigger clusters using specified measure of similarity between the pair of objects. The choice of which clusters to merge is determined by a linkage criteria. Thus, in agglomerative clustering, we start at the leaves and successively clusters are merged together to form the dendrogram. The clusters shall keep merging with each other until all of the objects are in a single cluster or until certain termination conditions are satisfied. The termination condition is decided by the researcher which depends upon the number of clusters required to be formed. One of the criteria in deciding the number of clusters to be formed depends upon whether some meanings can be attached to these clusters or not. Consider the following raw data in Fig. 10.2. Each data is a case/object and is considered to be the independent cluster. Depending upon the distances, these clusters are merged in different steps, and finally we get a single cluster. The formation of these clusters at different stages is shown in Fig. 10.3.

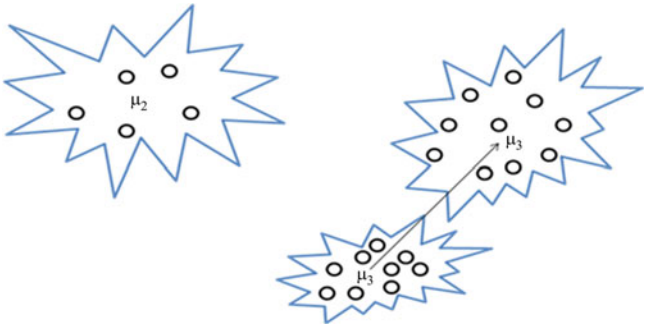


Fig. 10.4 Linkage of clusters using centroid method

In agglomerative clustering, different methods are used to form the clusters. These methods are discussed below.

Centroid Method

In this method, clusters are merged on the basis of the Euclidean distance between the cluster centroids. Clusters having least Euclidean distance between their centroids are merged together. In this method, if two unequal sized groups are merged together, then larger of the two tends to dominate the merged cluster. Since centroid methods compare the means of the two clusters, outliers affect it less than most other hierarchical clustering methods. However, it may not perform well in comparison to Ward's method or average linkage method (Milligan 1980). Linkage of clusters using centroid method is shown in Fig. 10.4.

Variance Methods

In this method, clusters are formed that minimize the within cluster variance. In other words, clusters are linked if the variation within the two clusters is least. This is done by checking the squared Euclidean distance to the center mean. The method used in checking the minimum variance in forming clusters is known as Ward's minimum variance method. This method tends to join the clusters having small number of observations and is biased towards producing clusters with same shape and with nearly equal number of observations. The variance method is very sensitive to the outliers. If "a" to "g" represents seven clusters then cluster formation using Ward's method can be shown graphically in Fig. 10.5.

Linkage Methods

In agglomerative clustering, clusters are formed on the basis of three different types of linkage methods.

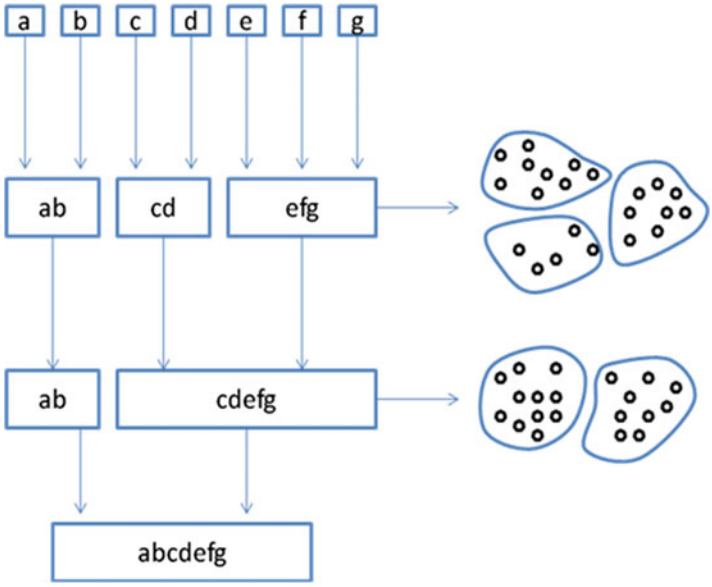
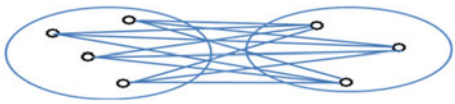


Fig. 10.5 Linkage of clusters using variance method

Fig. 10.6 Clusters based on single linkage

Fig. 10.7 Clusters based on complete linkage

Fig. 10.8 Clusters based on average linkage



1. *Single Linkage Method:* In this method, clusters are formed on the basis of minimum distance between the closest members of the two clusters. This is also known as nearest neighbor rule. This kind of linkage can be seen in Fig. 10.6.
2. *Complete Linkage Method:* In this method, clusters are formed on the basis of minimum distance between the farthest members of the two clusters. This is also known as furthest neighbor rule. Complete linkage can be shown by Fig. 10.7.
3. *Average Linkage Method*: This procedure uses the minimum average distance between all pairs of objects (in each pair one member must be from a different cluster) as the criteria to make the next higher cluster. Average linkage can be shown by Fig. 10.8.

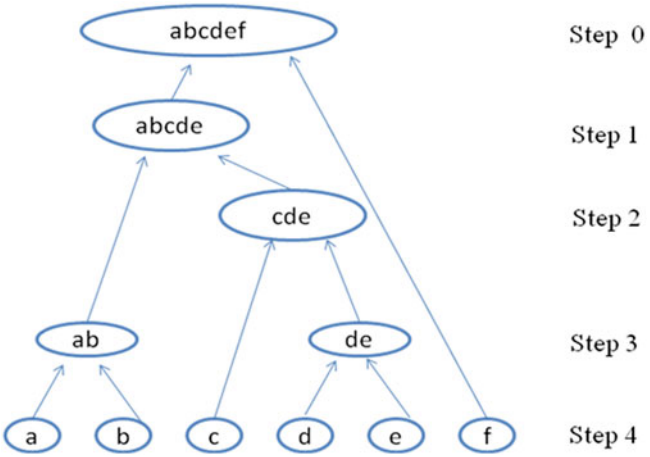


Fig. 10.9 Formation of clusters at different stages in divisive clustering

Divisive Clustering

In divisive clustering, we start by considering all individual objects/cases as one cluster (called as root) and recursively splits into smaller and smaller clusters owing to any of the distance criteria, until each object forms a cluster on its own or until it satisfies certain termination conditions which depend upon the number of clusters to be formed. Here, data objects are grouped in a top down fashion. Thus, in divisive clustering, we start at the root and reaches to leaves. Divisive clustering is just the reverse of agglomerative clustering. Cluster formation in divisive clustering schedule can be seen in Fig. 10.9.

Nonhierarchical Clustering (*K*-Means Cluster)

Unlike hierarchical clustering, in *K*-means clustering, a number of clusters are decided in advance. Solutions cannot be obtained for a range of clusters unless you rerun the analysis every time for different number of clusters. In *K*-means clustering, the first step is to find the *K*-centers. We start with an initial set of *K*-means and classify cases/objects based on their distances to the centers. Next, the cluster means are computed again using the cases/objects that are assigned to the cluster. After this, we reclassify all cases/objects based on the new set of means. This iterative process keeps going until cluster means do not change much between successive steps. Once the stability of cluster means is achieved, the means of the clusters are calculated once again and all the cases are assigned to their permanent clusters. If one can have a good guesses for the initial *K*-centers, those can be used as initial starting points; otherwise, let the SPSS find *K* cases that are well separated and use these values as initial cluster centers.

In hierarchical clustering, distance or similarity matrix between all pair of cases is required to be computed. This matrix becomes voluminous if the number of cases

is in thousands. Because of this, so much processing is required, and even with the modern computer, one needs to wait for some time to get the results. On the other hand, *k*-means clustering method does not require computation of all possible distances.

Nonhierarchical clustering solution has three different approaches, that is, sequential threshold method, parallel threshold, and optimizing partitioning method.

The *sequential threshold* method is based on finding a cluster center and then grouping all objects that are within a specified threshold distance from the center. Here, one cluster is created at a time.

In *parallel threshold* method, several cluster centers are determined simultaneously and then objects are grouped depending upon the specified threshold distance from these centers. These threshold distances may be adjusted to include more or fewer objects in the clusters.

The *optimizing partitioning* is similar to other two nonhierarchical methods except it allows for reassignment of objects to another cluster depending on some optimizing criterion. In this method, a nonhierarchical procedure is run first, and then objects are reassigned so as to optimize an overall criterion.

Precautions: *K*-means clustering is very sensitive toward the outliers because they will usually be selected as initial cluster centers. If outlier exists in the data, this will result in outliers forming clusters with small number of cases. Therefore, it is important for the researcher to screen the data for outliers and remove them before starting the cluster analysis.

Two-Step Cluster

Two-step clustering procedure is an exploratory statistical tools used for identifying the natural grouping of cases/objects within a large data set. It is an efficient clustering procedure in a situation where the data set is very large. This procedure has an ability to create clusters if some of the variables are continuous and others are categorical. It provides automatic identification of number of clusters present in the data.

There are two assumptions in this analysis: first, the variables are independent, and, second, each continuous variable follows a normal distribution whereas each categorical variable has a multinomial distribution. The two-step cluster analysis procedure provides solution in two steps which are explained as follows:

Step 1: Pre-cluster Formation

Pre-clusters are the clusters of original cases/objects that are used in place of raw data to reduce the size of the distance matrix between all possible pair of cases. After completing the pre-clustering, the cases in the same pre-cluster are treated as a single entity. Thus, the size of the distance matrix depends upon the number of pre-clusters instead of cases. Hierarchical clustering method is used on these pre-clusters instead of the original cases.

Step 2: Clustering Solutions Using Pre-clusters

In the second step, the standard hierarchical clustering algorithm is used on the pre-clusters for obtaining the cluster solution. The agglomerative clustering algorithm may be used to produce a range of cluster solutions. To determine which number of clusters is the best, each of these cluster solutions may be compared using either Schwarz's Bayesian criterion (BIC) or the Akaike information criterion (AIC) as the clustering criterion. The readers are advised to read about these procedures from some other texts.

Standardizing the Variables

Cluster analysis is normally used for the data measured on interval scale and rarely used for ratio data. In cluster analysis, distances are computed between the pair of cases on each of the variables. And if the units of measurement for these variables are different, then one must be worried about its impact on these distances.

Variables having larger values will have a larger impact on the distance compared to variables that have smaller values. In that case, one must standardize the variables to a mean of 0 and a standard deviation of 1.

If the variables are measured on interval scale and range of scale is same for each of the variable, then standardization of variables is not required, but if its range of measurement scale is different for different variables or if they are measured on ratio scale, then one must standardize the variables in some way so that they all contribute equally to the distance or similarity between cases.

Icicle Plots

It is the plotting of cases joining to form the clusters at each stage. You can see in Fig. 10.10 what is happening at each step of the cluster analysis when average linkage between groups is used to link the clusters. The figure is called an *icicle plot* because the columns representing cases look like icicles hanging from eaves. Each column represents one of the case/object you are clustering. Each row represents a cluster solution with different numbers of clusters.

If you look at the figure from bottom up, the last row (not shown) is the first step of the analysis. Each of the cases is a cluster of its own. The number of clusters at that point is 6. The five-cluster solution arises when the cases "a" and "b" are joined into a cluster. It is so because they had the smallest distance of all pairs. The four-cluster solution results from the merging of the cases "d" and "e" into a cluster. The three-cluster solution is the result of combining the cases "c" with "de." Going similarly, for the one cluster solution, all of the cases are combined into a single cluster.

		a		b		c		d		e		f	
Number of Clusters	1	X	X	X	X	X	X	X	X	X	X	X	X
	2	X	X	X	X	X	X	X	X	X	X		X
	3		X	X	X		X	X	X	X	X		X
	4		X	X	X		X		X	X	X		X
	5		X	X	X		X		X		X		X

Fig. 10.10 Vertical icicle plot

Remarks

1. When pairs of cases are tied for the smallest distance to form a cluster, an arbitrary selection is made. And, therefore, if cases are sorted differently, you might get a different cluster solution. But that should not bother you as there is no right or wrong answer to a cluster analysis. Many groupings are equally viable.
2. In case of large number of cases in cluster analysis, icicle plot can be developed by taking cases as rows. You must specify the "Horizontal" on the Cluster Plots dialog box.

The Dendrogram

The dendrogram is the graphical display of the distances on which clusters are combined. The dendrogram can be seen in Fig. 10.22 and is read from left to right. Vertical lines show joined clusters. The position of the line on the scale represents the distance at which clusters are joined. The observed distances are rescaled to fall into the range of 1–25, and hence you do not see the actual distances; however, the ratio of the rescaled distances within the dendrogram is the same as the ratio of the original distances. In fact, the dendrogram is the graphical representation of the information provided by the agglomeration schedule.

The Proximity Matrix

Consider the data of four employees on three different parameters age, income, and qualification as shown in the Table 10.4. Let us see how the proximity matrix is developed on these data.

The proximity matrix is the arrangement of squared Euclidean distances in rows and columns obtained between all pairs of cases. The squared Euclidean distances shall be computed by adding the squared differences between the two employees on each of the three variables.

Table 10.4 Employees' profile

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Employee 3	1.2	1.9	−0.9
Employee 4	1.5	−0.4	1.3

Table 10.5 Proximity matrix

	Squared Euclidean Distance			
Cases	Employee 1	Employee 2	Employee 3	Employee 4
Employee 1	0	0.38	12.83	10.05
Employee 2	0.38	0	2.25	7.25
Employee 3	12.83	2.25	0	10.22
Employee 4	10.05	7.25	10.22	0

The distance between employees 1 and 2 $= (2.5-2.3)^2 + (2.4-2.1)^2 + (2.4-1.9)^2 = .04 + .09 + .25 = 0.38$

The distance between employees 1 and 4 $= (2.5–1.5)^2 + (2.4 + 0.4)^2 + (2.4-1.3)^2 = 1.00 + 7.84 + 1.21 = 10.05$

The distance between employees 2 and 3 $= (2.3-1.2)^2 + (2.1-1.9)^2 + (1.9 + 0.9)^2 = 1.21 + 0.04 + 1.00 = 2.25$

This way, all distances can be computed which are shown in Table 10.5. This table is known as the proximity matrix.

All the entries in the diagonal are 0 because an employee does not differ with himself. The smallest difference between two employees is 0.38, the distance between the employee 1 and employee 2. The largest distance, 12.83, occurs between employee 1 and employee 3. The distance matrix is symmetric, and, therefore, you can see that the distance between the first and third employee is same as the distance between the third and first employee.

What We Do in Cluster Analysis

In using cluster analysis, one needs to follow different steps to get the final results. You may not understand all the steps at this moment but use it as a blueprint of the analysis and proceed further, and I am sure by the time you finish reading the entire chapter, you will have a fairly good idea about its application. Once you understand different concepts of cluster analysis discussed in this chapter, you will be taken to a solved example by using SPSS, and this will give you practical knowledge of using this analysis to your data set with SPSS. Below are the steps which are used in cluster analysis:

1. Identify the variables on which subjects/objects need to be clustered.

2. Select the distance measure for computing distance between cases. One can choose any of the distance measures like squared Euclidean distance, Manhattan distance, Chebyshev distance, or Mahalanobis (or correlation) distance.
3. Decide the clustering procedure to be used from the wide variety of clustering procedure available in the hierarchical or nonhierarchical clustering sections.
4. Decide on the number of clusters to be formed. The sole criteria in deciding the number of clusters is based on the fact that one should be able to explain these clusters on the basis of their characteristics.
5. Map and interpret clusters using illustrative techniques like perceptual maps, icicle plots, and dendrograms and draw conclusions.
6. Assess reliability and validity of the obtained clusters by using any one or more of the following methods:

   (i) Apply the cluster analysis on the same data by using different distance measure.
   (ii) Apply the cluster analysis on the same data by using different clustering technique.
   (iii) Split the same data randomly into two halves and apply the cluster analysis separately on each part.
   (iv) Repeat cluster analysis on same data several times by deleting one variable each time.
   (v) Repeat cluster analysis several times, using a different order each time.

Assumptions in Cluster Analysis

Following assumptions need to be satisfied in cluster analysis:

1. The cluster analysis is usually used for the data measured on interval scale. However, it can be applied for any type of data. If the variable set includes continuous as well as categorical, then two-step cluster should be used.
2. The variables in the cluster analysis should be independent with each other.
3. Inter-object similarity is often measured by Euclidean distance between pairs of objects.
4. The data needs to be standardized if the range or scale of measurement of one variable is much larger or different from the range of others.
5. In case of nonstandardized data, Mahalanobis distance is preferred as it compensates for intercorrelation among the variables.
6. In applying two-step cluster with continuous as well as categorical variables, it is assumed that the continuous variables are normally distributed whereas categorical variables have multinomial distribution.

Research Situations for Cluster Analysis Application

Cluster analysis can be applied to a wide variety of research problems in the area of management, psychology, medicine, pharmaceuticals, social sciences, etc. Following are the situations where this technique can be applied:

1. Cluster analysis can be used to classify the consumer population into market segments for understanding the requirements of potential customers in different groups. Such studies may be useful in segmenting the market, identifying the target market, product positioning, and developing new products.
2. In a big departmental store, all inventories may be clustered into different groups for placing them in same location or giving the similar code for enhancing sale and easy monitoring of the products.
3. In the field of psychiatry, the cluster analysis may provide the cluster of symptoms such as paranoia and schizophrenia, which is essential for successful therapy.
4. In educational research, all schools of a district can be classified into different clusters on the basis of the parameters like number of children, teacher's strength, total grant, school area, and location to develop and implement the programs and policies effectively for each of these groups separately.
5. In the area of mass communication, television channels may be classified into homogenous groups based on certain characteristics like TRP, number of programs televised per week, number of artists engaged, coverage time, programs in different sectors, advertisements received, and turnover. Different policies may be developed for different groups of channels by the regulatory body.
6. In medical research, cluster analysis may provide the solution for clustering of diseases so that new drugs may be developed for different clusters of diseases. This analysis may also be useful in clustering the patients on the basis of symptoms for easy monitoring of drug therapy on mass scale.

Steps in Cluster Analysis

By learning terminologies involved in cluster analysis and the guidelines discussed in the heading "What We Do in Cluster Analysis?", you are now in a better position to understand the procedure of its use for addressing your objectives. The cluster analysis is usually done in two stages. The whole analysis is carried out in two stages, the details of which have been discussed in the following steps:

Stage 1

1. The first step in cluster analysis is to apply the hierarchical cluster analysis in SPSS to find the agglomerative schedule and proximity matrix for the data obtained on each of the variables for all the cases. To form clusters, you need

to select a criterion for determining similarity or distance between cases and a linkage criterion for merging clusters at successive steps. After doing so, the SPSS output provides proximity matrix which shows the distances (similarity) between all the cases/objects and agglomerative schedule which is used to find the number of clusters present in the data on the basis of fusion coefficients. The detailed discussion as to how to do it shall be made while discussing the solved example of cluster analysis using SPSS.

2. Prepare icicle plot and dendrogram of the data. These two figures can be obtained by providing options in the SPSS. The icicle plot is the visual representation of the agglomerative schedule whereas the dendrogram plot shows how distant (or close) cases are when they are combined.

Stage 2

3. The second step in cluster analysis is to apply the *K*-means cluster analysis in SPSS. The process is not stopped in the first stage just because of the fact that *K*-means analysis provides much stable clusters due to interactive procedure involved in it in comparison to the single-pass hierarchical methods. The *K*-means analysis provides four outputs, namely, initial cluster centers, case listing of cluster membership, final cluster centers, and analysis of variance for all the variables in each of the clusters.
4. The case listing of cluster membership is used to describe as to which case belongs to which of the clusters.
5. The final cluster centers are obtained by doing iteration on the initial cluster solutions. It provides the final solution. On the basis of final cluster centers, the characteristics of different clusters are explained.
6. Finally, ANOVA table describes as to which of the variables is significantly different across all the identified clusters in the problem.

The detailed discussion of the above-mentioned outputs in cluster analysis shall be done by means of the results obtained in the solved example using SPSS.

Solved Example of Cluster Analysis Using SPSS

Example 10.1 A media company wants to cluster its target audience in terms of their preferences toward quality, contents, and features of FM radio stations. Twenty randomly chosen students were selected from a university who served the sample for the study. Below-mentioned 14 questions were finally selected by their research team after the content and item analysis which measured many of the variables of interest. The respondents were asked to mark their responses on a 5-point scale where 1 represented complete disagreement and 5 complete agreement. The responses of the respondents on all the 12 questions that measured different dimensions of FM stations are shown in Table 10.6.

Table 10.6 Response of students on the questions related to quality, contents, and features of FM radio stations

SN	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
1	5	4	2	5	1	3	5	2	1	4	3	4	3	4
2	1	4	4	2	5	2	3	5	2	3	4	2	2	3
3	2	2	3	3	2	4	2	3	4	4	2	2	4	4
4	5	3	3	4	4	4	5	3	2	5	2	5	3	5
5	4	1	2	4	1	1	5	4	2	4	3	4	2	4
6	4	2	3	4	2	5	2	1	5	2	1	3	5	3
7	2	3	2	2	3	4	3	4	4	3	4	3	5	2
8	5	2	2	5	2	4	5	2	2	4	1	5	1	4
9	2	4	4	2	5	3	4	4	3	3	5	2	2	2
10	3	4	4	2	4	3	2	5	2	4	3	3	2	2
11	4	5	4	3	5	4	4	4	1	1	5	4	3	2
12	2	4	5	1	4	2	2	4	2	4	3	4	4	3
13	2	5	4	2	5	3	3	5	3	2	5	3	3	2
14	1	5	4	5	4	3	2	5	3	3	5	4	4	2
15	2	5	5	3	4	2	3	4	4	3	4	3	3	3
16	5	3	2	4	5	2	4	4	3	5	2	5	2	5
17	5	2	3	5	2	3	5	2	4	5	3	4	4	4
18	5	2	2	2	2	4	4	3	4	2	2	2	4	1
19	4	3	3	3	4	5	2	3	5	4	3	2	5	2
20	3	4	4	1	2	4	4	2	4	2	3	4	4	3

Questions on quality, contents, and features of FM stations

1. The FM station should provide more old Hindi songs.
2. FM stations must help an individual in solving their personal problems.
3. The presentation style of RJs helps popularizing an FM station.
4. An FM station should provide some kind of prizes/incentives to its listeners.
5. The station must telecast latest songs more often.
6. The FM stations must contain more entertaining programs.
7. Popularity of RJs depends upon their humor and ability to make program interesting.
8. FM station should provide more opportunity to listeners to talk to celebrities.
9. RJs voice must be clear and melodious.
10. FM channels should play 24×7.
11. FM stations should give information for other sports along with cricket.
12. FM stations should provide information regarding educational/professional courses available in the city.
13. FM stations should provide information regarding different shopping offers available in the city.
14. RJs should speak in an understandable language, preferably in local language.

Solution In earlier chapters, you have seen the procedure of applying different statistical techniques by using SPSS. By now, you must have been well acquainted with the procedure of starting the SPSS on the system, defining variables and their

characteristics and preparing data file, and, therefore, these steps shall be skipped in this chapter. In case of any clarification, readers are advised to go through Chap. 1 for detailed guidelines for preparing the data file.

The steps involved in using SPSS for cluster analysis shall be discussed first, and then the output obtained from the analysis shall be shown and explained. The whole scheme of cluster analysis with SPSS is as follows:

Stage 1

First of all, the hierarchical cluster analysis shall be done by using the sequence of SPSS commands. The following outputs would be generated in this analysis:

(a) Proximity matrix of distances (similarity) between all the cases/objects
(b) Agglomerative schedule
(c) Icicle plot
(d) Dendrogram

On the basis of fusion coefficients in the agglomerative schedule, the number of clusters (say K) is decided.

Stage 2

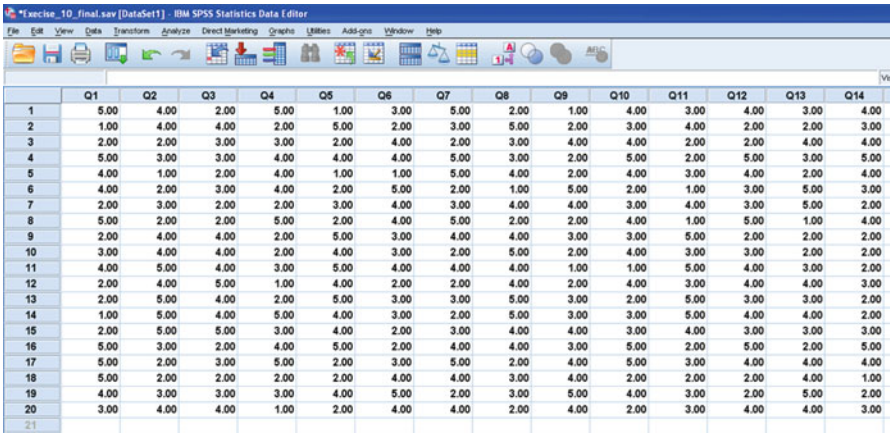
After deciding the number of clusters in the hierarchical cluster analysis, the data is again subjected to K-means cluster analysis in SPSS. Using this analysis, the following outputs would be generated:

(a) Initial cluster centers
(b) Case listing of cluster membership
(c) Final cluster centers
(d) Analysis of variance for comparing the clusters on each of the variables

Stage 1: SPSS Commands for Hierarchical Cluster Analysis

(a) ***Data file*** After defining variable names and their labels, prepare the data file for the responses of the students on all the variables shown in Table 10.2. The data file shall look like as shown in Fig. 10.11.
(b) ***Initiating command for hierarchical cluster analysis:*** After preparing the data file, start the hierarchical analysis in SPSS by the following command sequence (Fig. 10.12):

Analyze → Classify → Hierarchical Cluster



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
1	5.00	4.00	2.00	5.00	1.00	3.00	5.00	2.00	1.00	4.00	3.00	4.00	3.00	4.00
2	1.00	4.00	4.00	2.00	5.00	2.00	3.00	5.00	2.00	3.00	4.00	2.00	2.00	3.00
3	2.00	2.00	3.00	3.00	2.00	4.00	2.00	3.00	4.00	4.00	2.00	2.00	4.00	4.00
4	5.00	3.00	3.00	4.00	4.00	4.00	5.00	3.00	2.00	5.00	2.00	5.00	3.00	5.00
5	4.00	1.00	2.00	4.00	1.00	1.00	5.00	4.00	2.00	4.00	3.00	4.00	2.00	4.00
6	4.00	2.00	3.00	4.00	2.00	5.00	2.00	1.00	5.00	2.00	1.00	3.00	5.00	3.00
7	2.00	3.00	2.00	2.00	3.00	4.00	3.00	4.00	4.00	3.00	4.00	3.00	5.00	2.00
8	5.00	2.00	2.00	5.00	2.00	4.00	5.00	2.00	2.00	4.00	1.00	5.00	1.00	4.00
9	2.00	4.00	4.00	2.00	5.00	3.00	4.00	4.00	3.00	3.00	5.00	2.00	2.00	2.00
10	3.00	4.00	4.00	2.00	4.00	3.00	2.00	5.00	2.00	4.00	3.00	3.00	2.00	2.00
11	4.00	5.00	4.00	3.00	5.00	4.00	4.00	4.00	1.00	1.00	5.00	4.00	3.00	2.00
12	2.00	4.00	5.00	1.00	4.00	2.00	2.00	4.00	2.00	4.00	3.00	4.00	4.00	3.00
13	2.00	5.00	4.00	2.00	5.00	3.00	3.00	5.00	3.00	2.00	5.00	3.00	3.00	2.00
14	1.00	5.00	4.00	5.00	4.00	3.00	2.00	5.00	3.00	3.00	5.00	4.00	4.00	2.00
15	2.00	5.00	5.00	3.00	4.00	2.00	3.00	4.00	4.00	3.00	4.00	3.00	3.00	3.00
16	5.00	3.00	2.00	4.00	5.00	2.00	4.00	4.00	3.00	5.00	2.00	5.00	2.00	5.00
17	5.00	2.00	3.00	5.00	2.00	3.00	5.00	2.00	4.00	5.00	3.00	4.00	4.00	4.00
18	5.00	2.00	2.00	2.00	2.00	4.00	4.00	3.00	4.00	2.00	2.00	2.00	4.00	1.00
19	4.00	3.00	3.00	3.00	4.00	5.00	2.00	3.00	5.00	4.00	3.00	2.00	5.00	2.00
20	3.00	4.00	4.00	1.00	2.00	4.00	4.00	2.00	4.00	2.00	3.00	4.00	4.00	3.00
21														

Fig. 10.11 Showing data file for all the variables in SPSS

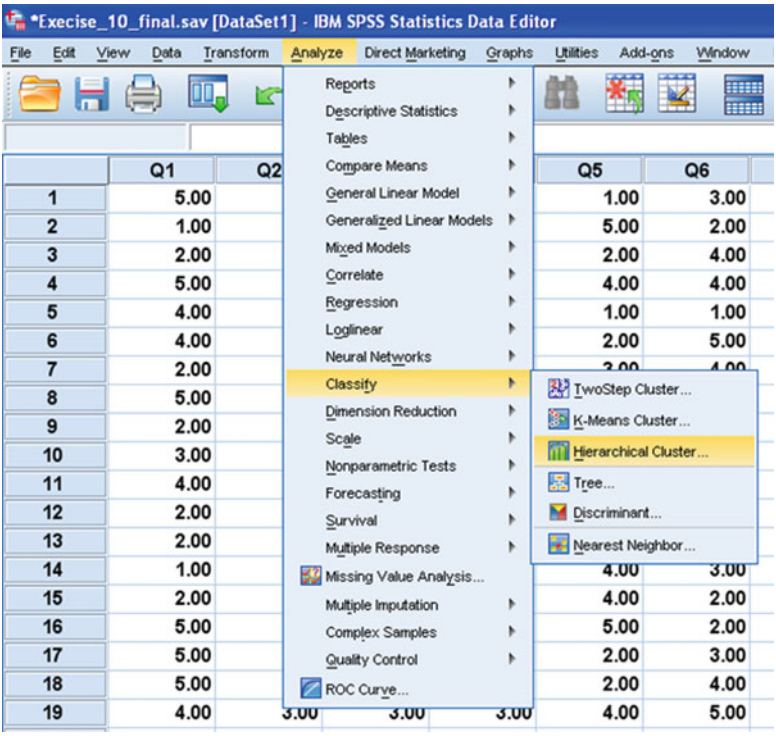


Fig. 10.12 Sequence of SPSS commands for hierarchical cluster

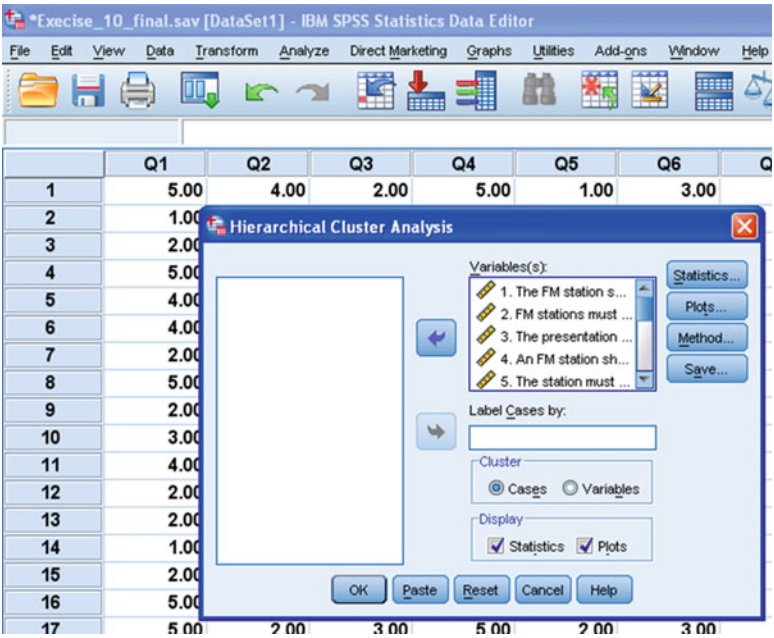


Fig. 10.13 Selecting variables for hierarchical analysis

(i) *Selecting variables for analysis:* After clicking the **Hierarchical Cluster** option, you will be taken to the next screen for selecting variables. Select the variables as follows:

- Select all the variables and bring them in the "Variable(s)" section.
- Ensure that in the "Display" section, the options "Statistics" and "Plots" are checked. These are selected by default.
- In case if a variable denoting label of each cases is defined in the variable view while preparing the data file, then bring that variable under the section "Label Cases by." While defining the variable for label in the variable view, define its variable type as String under the column heading **Type.** However, for the time being, you can skip the process of defining the variable for label and leave the option "Label Cases by" blank.
  The screen will look like as shown in Fig. 10.13.

(ii) *Selecting options for computation:* After selecting the variables, you need to define different options for generating all the four outputs of hierarchical analysis. Take the following steps:

- Click the tag **Statistics** in the screen shown in Fig. 10.13 and take the following steps:
- Ensure that the "Agglomerative schedule" is checked. By default, it is checked.

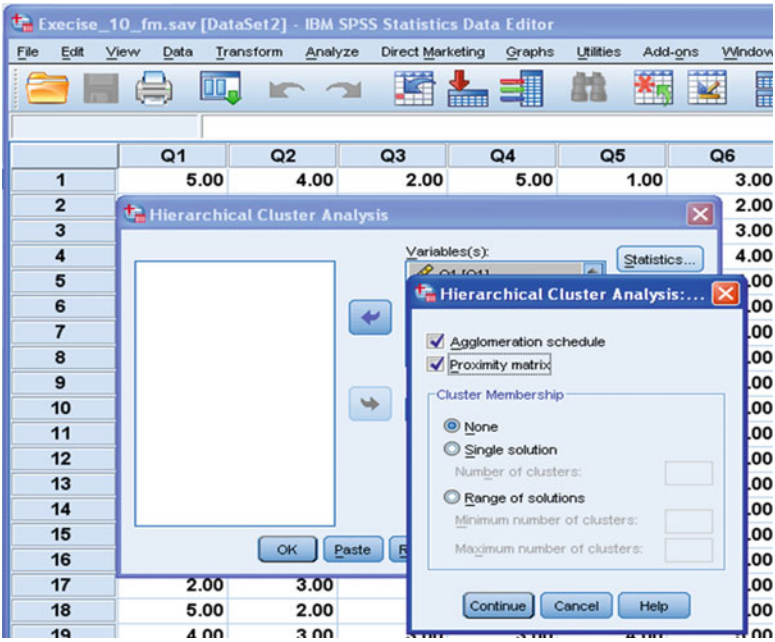


Fig. 10.14 Screen showing option for generating agglomerative schedule and proximity matrix

- Check "Proximity matrix."
- Leave other options by default and click ***Continue***.

The screen will look like Fig. 10.14.

- Click the tag **Plots** in the screen shown in Fig. 10.13 and take the following steps:
  - Check the option "Dendrogram."
  - Ensure that the option "All clusters" is checked in the "icicle plot" section. This is checked by default.
  - Ensure that the option "Vertical" is checked in the "Orientation section." This is also checked by default. The option "Vertical" is selected if the number of cases is small. However, if the number of cases is large, then select "Horizontal."
  - Click ***Continue***.

The screen will look like Fig. 10.15.

- Click the tag **Method** in the screen shown in Fig. 10.13 and do the following steps:
  - Select the option "Ward's method" as cluster method. You can choose any other linkage method. For details read the methods under the heading

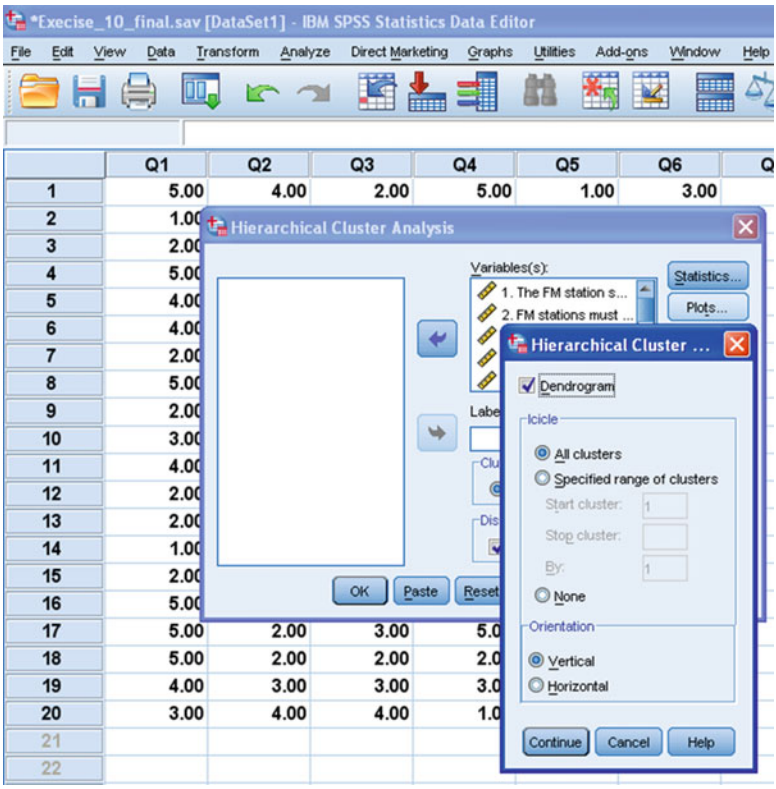


Fig. 10.15 Selecting options for dendrogram and icicle plot

Distance Method discussed earlier. Personally I prefer the Ward's method as it depends upon the minimum variance concept and gives the clusters which are more homogenous within itself.

- Select the option "Squared Euclidean distance" as an interval measure. However, you can choose any other method as a distance measure like Euclidean distance, Pearson correlation method, or Chebyshev method. But generally squared Euclidean method is used to find the distance in the proximity matrix.
- Select the option "None" in the "Transform Values" section. This is so because in our example, the units of measurement for all the variables are same. However, if the units of measurements are different, one needs to standardize the variables. The most popular transformation is "Z-scores" which needs to be selected if the measurement units are different for all the variables.
- Check the option "Absolute values" in the "Transform Measures" option. This option transforms the values generated by the distance measure. This option is not required for squared Euclidean distance.

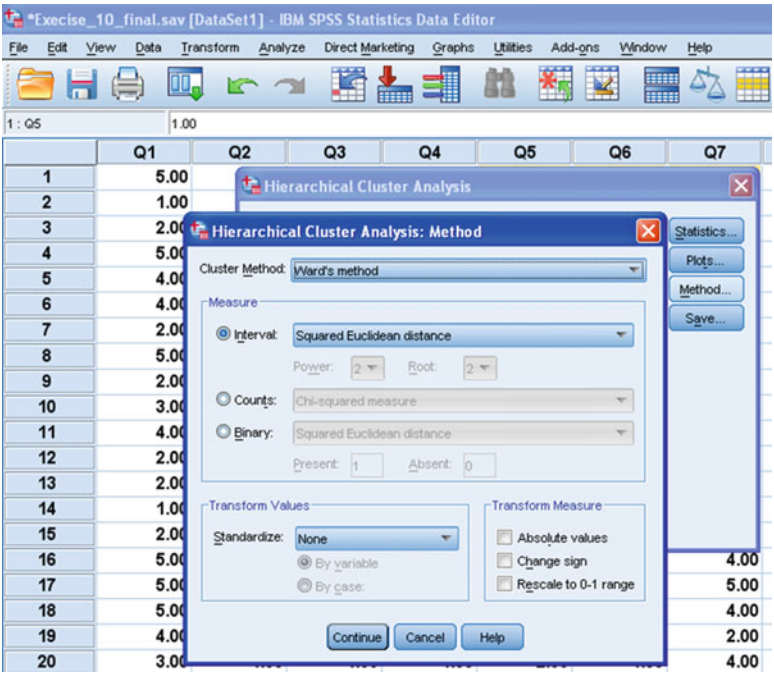


Fig. 10.16 Selecting options for cluster method and distance measure criteria

- Click Continue. You will be taken back to the screen shown in Fig. 10.13. The screen will look like as shown in Fig. 10.16.

- Click **OK**

(c) ***Getting the output:*** Clicking the option **OK** shall generate lot of outputs in the output window. The four outputs that would be selected are Proximity matrix, Agglomerative schedule, Icicle plot, and Dendrogram. These outputs have been shown in Tables 10.7, 10.8 and Fig. 10.21, 10.22.

Stage 2: SPSS Commands for K*-Means Cluster Analysis*

Stage 1 was the explorative process where number of initial clusters was identified. These initial clusters were identified on the basis of fusion coefficients in the agglomerative schedule. After deciding the number of clusters, apply the *K*-means cluster analysis in stage 2. In stage 1, three clusters were identified on the basis of the agglomeration schedule in Table 10.8 (for details, see Interpretation of Findings). This shall be used to find the final solution in the *K*-means cluster

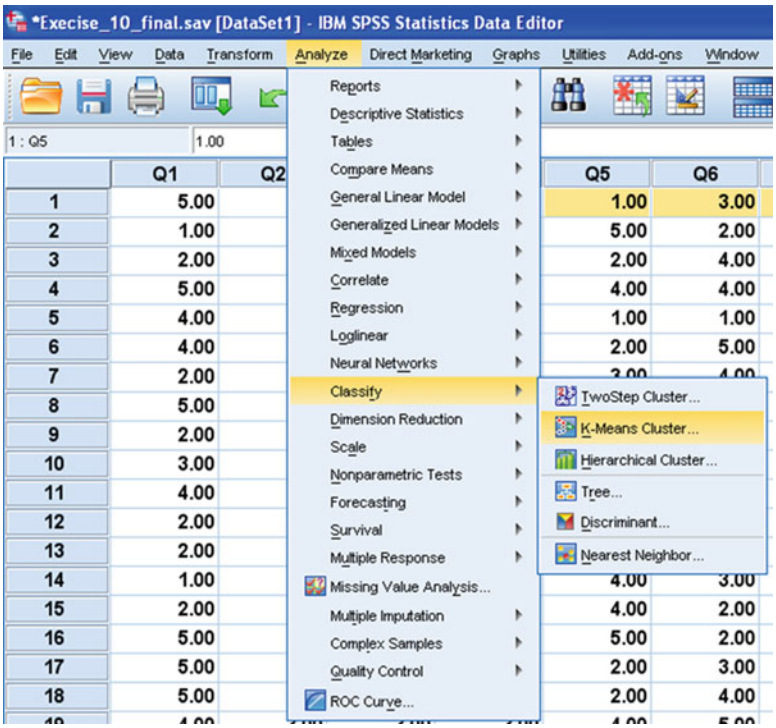


Fig. 10.17 Sequence of SPSS commands for *K*-means cluster

analysis. The data file developed for the hierarchical analysis is also used for the *K*-means cluster analysis. Follow these steps in stage 2.

(i) *Initiating command for K-means cluster analysis:* Start the *K*-means analysis by using the following command sequence (Fig. 10.17):

Analyze → Classify → K-Means Cluster Analysis

(ii) *Selecting variables for analysis:* After clicking the ***K*-Means Cluster Analysis** option, you will be taken to the next screen for selecting variables. Select the variables as follows:

- Select all the variables and bring them in the "Variable(s)" section.
- Write number of clusters as 3. This is so because only three clusters were identified from the hierarchical analysis.
- Click the option **Iterate** and ensure that the minimum iteration is written as 10. In fact, this is done by default. If you want to have more than 10 maximum iterations, it may be mentioned here.
- Click ***Continue.***

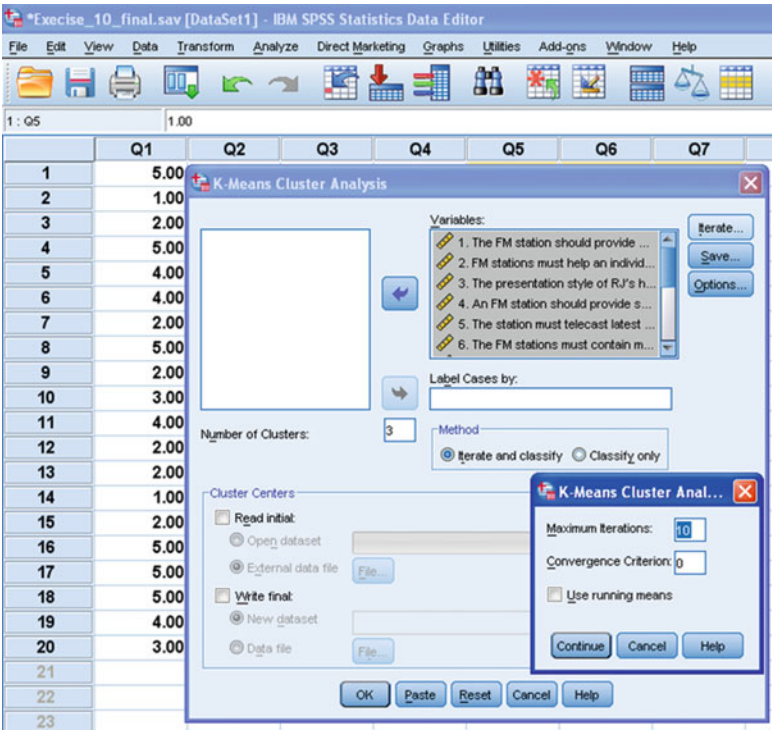


Fig. 10.18 Screen showing selection of variables for *K*-means analysis

The screen shall look like Fig. 10.18.

– Click the tag **Save** and take the following steps:

  – Check the option "Cluster membership."
  – Check the option "Distance from cluster center."

– Click ***Continue.***

The screen shall look like Fig. 10.19.

– Click the tag **Options** and take the following steps:

  – Ensure that the option "Initial cluster centers" is checked. In fact, this is checked by default.
  – Check the option "ANOVA table."
  – Check the option "Cluster information for each case."

– Click ***Continue.***

  The screen would look like Fig. 10.20.

– Click **OK**.

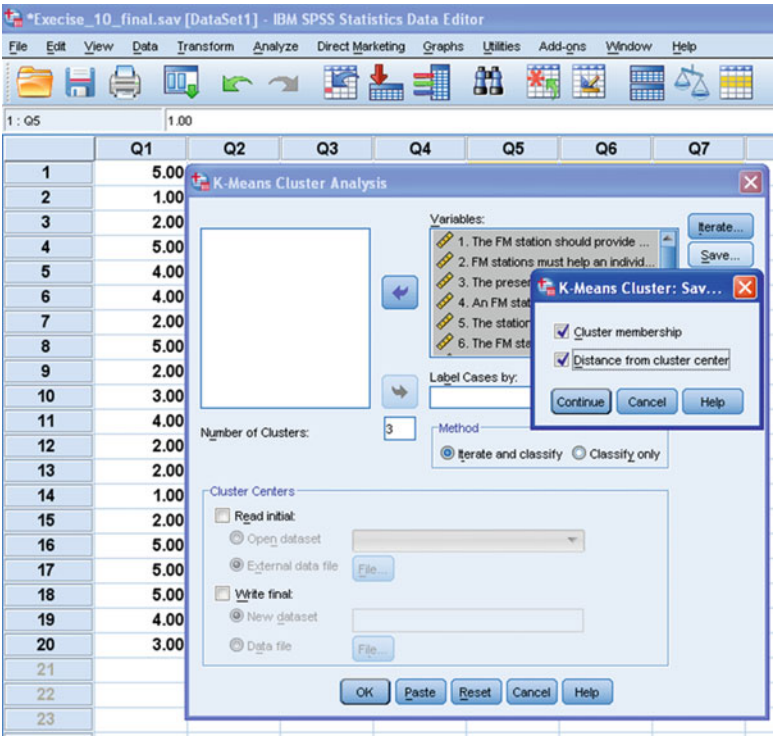


Fig. 10.19 Screen showing option for getting cluster memberships and distance from cluster center

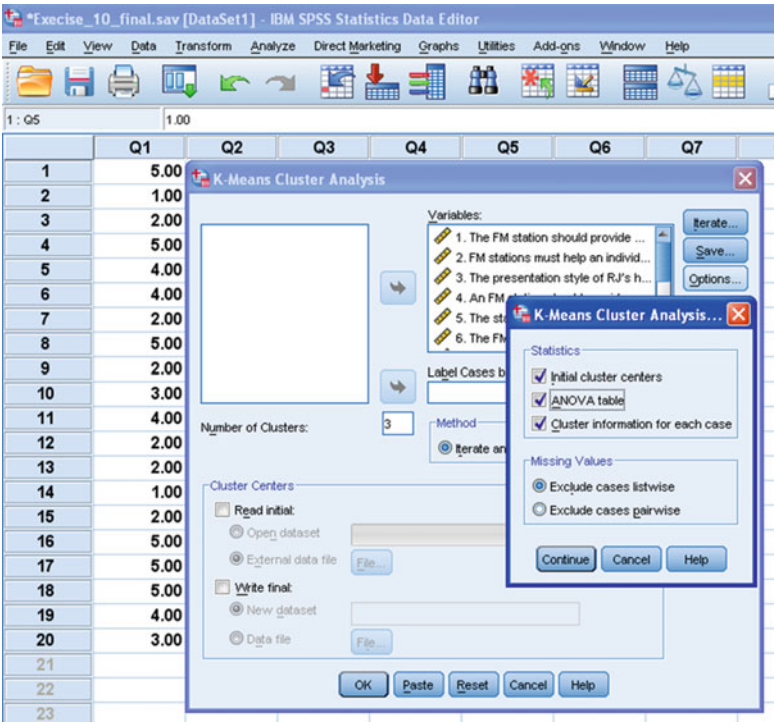


Fig. 10.20 Screen showing options for cluster information and ANOVA

Interpretations of Findings

Stage 1: The agglomerative cluster analysis done in stage 1 provided the outputs shown in Tables 10.7 and 10.8 and in Figs. 10.21 and 10.22. The agglomerative analysis is explorative in nature. Its primary purpose is to identify the initial cluster solution. Therefore, one should take all possible parameters to identify the clusters so that important parameters are not left out. We shall now discuss the results generated in the agglomerative analysis in stage 1.

Proximity Matrix: To Know How Alike (or Different) the Cases Are

Table 10.7 is a proximity matrix which shows distances between the cases. One can choose any distance criterion like squared Euclidean distance, Manhattan distance, Chebyshev distance, Mahalanobis (or correlation) distance, or Pearson correlation distance. In this example, the squared Euclidean distance was chosen as a measure of distance. The minimum distance exists between the 9th and 13th cases which is 6.00, whereas the maximum distance is observed between the 8th and 13th cases which is 87.00. The minimum distance means that these two cases would combine at the very first instance. This can be seen from Table 10.8 where 9th and 13th cases are combined into a single cluster in the very first stage. Similarly, the 8th and 13th cases are in the extreme clusters which can be seen in the dendrogram shown in Fig. 10.22.

Agglomerative Schedule: To Know How Should Clusters Be Combined

Table 10.8 is an agglomerative schedule which shows how and when the clusters are combined. The agglomerative schedule is used to decide the number of clusters present in the data and one should identify the number of clusters by using the column labeled “Coefficients” in this table. These coefficients are also known as fusion coefficients. The values under this column are the distance (or similarity) statistic used to form the cluster. From these values, you get an idea as to how the clusters have been combined. In case of using dissimilarity measures, small coefficients indicate that those fairly homogenous clusters are being attached to each other. On the other hand, large coefficients show that the dissimilar clusters are being combined. In using similarity measures, the reverse is true, that is, large coefficients indicate that the homogeneous clusters are being attached to each other, whereas small coefficients reveal that dissimilar clusters are being combined.

The value of fusion coefficient depends on the clustering method and the distance measure you choose. These coefficients help you decide how many clusters you need to represent the data. The process of cluster formation is stopped when the increase (for distance measures) or decrease (for similarity measures) in

Table 10.7 Proximity matrix

	Squared Euclidean distance																			
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.000	68.000	45.000	19.000	21.000	52.000	52.000	16.000	61.000	51.000	49.000	60.000	69.000	62.000	54.000	33.000	17.000	45.000	58.000	42.000
2	68.000	.000	39.000	57.000	55.000	82.000	30.000	80.000	7.000	11.000	29.000	16.000	9.000	24.000	12.000	49.000	71.000	59.000	48.000	42.000
3	45.000	39.000	.000	40.000	40.000	19.000	17.000	47.000	40.000	30.000	64.000	31.000	46.000	43.000	31.000	46.000	30.000	28.000	17.000	25.000
4	19.000	57.000	40.000	.000	30.000	51.000	49.000	15.000	52.000	40.000	46.000	45.000	60.000	61.000	47.000	10.000	16.000	50.000	45.000	41.000
5	21.000	55.000	40.000	30.000	.000	65.000	49.000	23.000	56.000	46.000	66.000	55.000	68.000	67.000	53.000	28.000	22.000	44.000	67.000	51.000
6	52.000	82.000	19.000	51.000	65.000	.000	34.000	48.000	71.000	61.000	73.000	62.000	73.000	68.000	56.000	65.000	33.000	23.000	20.000	28.000
7	52.000	30.000	17.000	49.000	49.000	34.000	.000	66.000	23.000	25.000	37.000	26.000	21.000	26.000	24.000	53.000	41.000	21.000	14.000	18.000
8	16.000	80.000	47.000	15.000	23.000	48.000	66.000	.000	73.000	57.000	67.000	74.000	87.000	84.000	70.000	25.000	21.000	47.000	64.000	52.000
9	61.000	7.000	40.000	52.000	56.000	71.000	23.000	73.000	.000	14.000	20.000	23.000	6.000	25.000	11.000	50.000	58.000	44.000	35.000	31.000
10	51.000	11.000	30.000	40.000	46.000	61.000	25.000	57.000	14.000	.000	26.000	11.000	14.000	25.000	15.000	36.000	54.000	40.000	31.000	33.000
11	49.000	29.000	64.000	46.000	66.000	73.000	37.000	67.000	20.000	26.000	.000	35.000	14.000	29.000	27.000	54.000	64.000	50.000	49.000	35.000
12	60.000	16.000	31.000	45.000	55.000	62.000	26.000	74.000	23.000	11.000	35.000	.000	19.000	28.000	14.000	45.000	57.000	53.000	38.000	26.000
13	69.000	9.000	46.000	60.000	68.000	73.000	21.000	87.000	6.000	14.000	14.000	19.000	.000	15.000	9.000	56.000	70.000	50.000	37.000	31.000
14	62.000	24.000	43.000	61.000	67.000	68.000	26.000	84.000	25.000	25.000	29.000	28.000	15.000	.000	14.000	59.000	61.000	67.000	40.000	46.000
15	54.000	12.000	31.000	47.000	53.000	56.000	24.000	70.000	11.000	15.000	27.000	14.000	9.000	14.000	.000	43.000	47.000	49.000	32.000	24.000
16	33.000	49.000	46.000	10.000	28.000	65.000	53.000	25.000	50.000	36.000	54.000	45.000	56.000	59.000	43.000	.000	26.000	58.000	51.000	55.000
17	17.000	71.000	30.000	16.000	22.000	33.000	41.000	21.000	58.000	54.000	64.000	57.000	70.000	61.000	47.000	26.000	.000	36.000	35.000	37.000
18	45.000	59.000	28.000	50.000	44.000	23.000	21.000	47.000	44.000	40.000	50.000	53.000	50.000	67.000	49.000	58.000	36.000	.000	21.000	23.000
19	58.000	48.000	17.000	45.000	67.000	20.000	14.000	64.000	35.000	31.000	49.000	38.000	37.000	40.000	32.000	51.000	35.000	21.000	.000	28.000
20	42.000	42.000	25.000	41.000	51.000	28.000	18.000	52.000	31.000	33.000	35.000	26.000	31.000	46.000	24.000	55.000	37.000	23.000	28.000	.000

This is a dissimilarity matrix

Table 10.8 Agglomeration schedule

	Cluster combined			Stage cluster first appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next stage
1	9	13	3.000	0	0	2
2	2	9	7.333	0	1	5
3	4	16	12.333	0	0	17
4	10	12	17.833	0	0	12
5	2	15	24.000	2	0	12
6	7	19	31.000	0	0	8
7	1	8	39.000	0	0	9
8	3	7	48.000	0	6	15
9	1	17	58.000	7	0	11
10	18	20	69.500	0	0	13
11	1	5	81.500	9	0	17
12	2	10	94.333	5	4	14
13	6	18	107.500	0	10	15
14	2	14	121.667	12	0	16
15	3	6	137.000	8	13	18
16	2	11	153.750	14	0	18
17	1	4	172.417	11	3	19
18	2	3	261.524	16	15	19
19	1	2	389.800	17	18	0

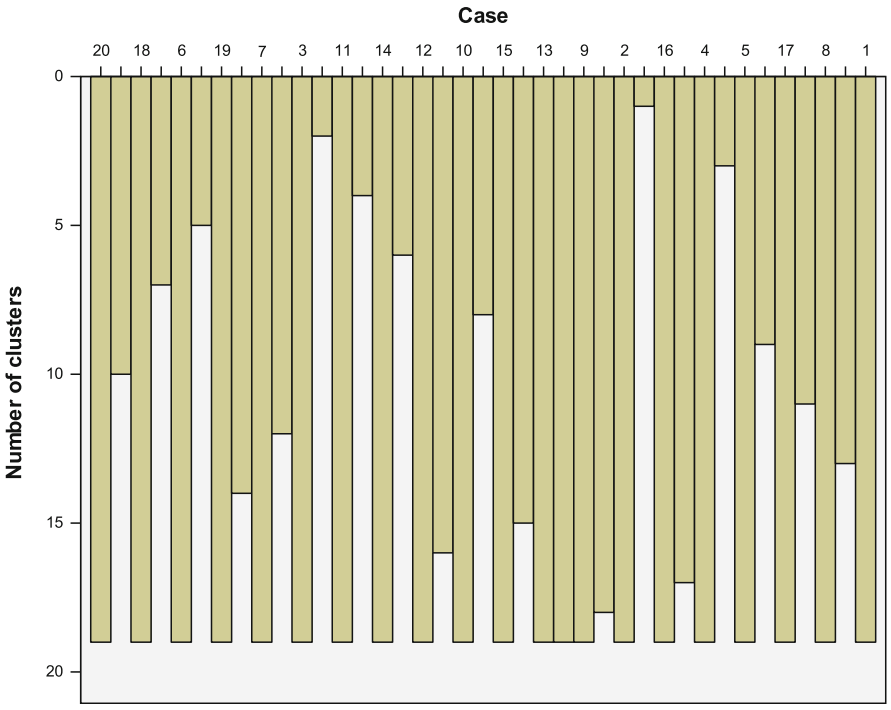


Fig. 10.21 Vertical icicle plot using Ward's linkage

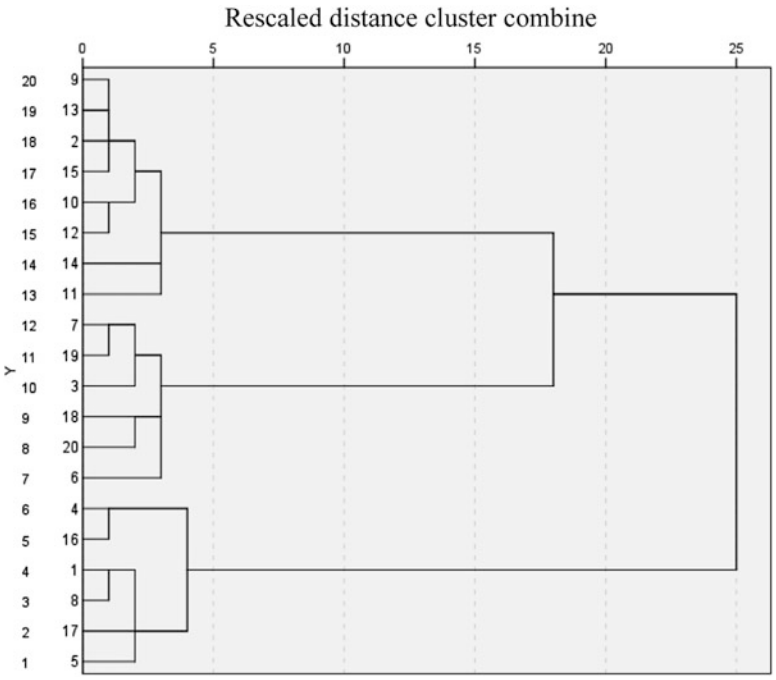


Fig. 10.22 Dendrogram using Ward linkage

the coefficients between the two adjacent steps is large. In this example, the process can be stopped at the three cluster solution, after stage 17. Let us see how it is done?

We should look for the coefficients from the last row upward because we want the lowest possible number of clusters due to economy and its interpretability. Stage 20 represents a one cluster solution where all the cases are combined into one cluster, and, therefore, it is not shown in Table 10.8. The largest difference (389.800–261.524) exists in the coefficients between stages 18 and 19, which means we have to stop the process of cluster formation after stage 19; this would result in only two-cluster solution. However, we may not be interested to represent the data by two clusters only; therefore, we will look for the next larger difference of (261.524–172.417) which is equal to 89.107 (between stage 18, the three-cluster solution, and stage 17, the four-cluster solution). The next one after that is (172.417–153.750), only 18.667, between stages 17 and 16. Thereafter, the difference keeps decreasing. So we decide to stop the cluster formation at stage 18 which is a three-cluster solution.

Thus, in general, the strategy is to first identify the largest difference in the coefficients and identify the stage of the lowest coefficient as the cluster solution. However, it is up to the researcher to decide the number of clusters depending upon its interpretability. You can see from the dendrogram shown in Fig. 10.22 that three clusters are clearly visible in this case.

The agglomeration schedule starts off using the case numbers that has smallest distance as shown by the icicle plot in Fig. 10.21. The cluster is formed by adding cases. The number of the lowest case becomes the number of this newly formed cluster. For example, if a cluster is formed by merging cases 3 and 6, it would be known as cluster 3, and if the clusters are formed by merging cases 3 and 1, then it would be known as cluster 1.

The columns labeled "Stage Cluster First Appears" shows the step at which each of the two clusters that are being joined first appear. For example, at stage 9 when clusters 1 and 17 are combined, it tells you that cluster 1 was first formed at stage 7 and cluster 17 is a single case, and that the resulting cluster (known as 1) will see action again at stage 11 (under the column "Next stage"). If number of cases are small then the icicle plot explains step-by-step clustering summary better than the agglomeration schedule.

The Icicle Plot: Summarizing the Steps

Figure 10.21 is the icicle plot which is a graphical representation of agglomerative schedule. It tells you how the clusters are formed at each stage. The figure is called an icicle plot because the columns look like icicles hanging from eaves. Each column represents one of the objects you are clustering. Each row shows a cluster solution with different number of clusters. You see the figure from the bottom up. The last row (not shown) is the first step of the analysis. Each of the cases is a cluster of itself. The number of clusters at this point is 20. The nineteen-cluster solution arises when cases 9 and 13 are joined into a cluster. It happened because they had the smallest distance among all pairs. The eighteen-cluster solution results from the merging of case 2 with cluster 9 into a cluster. This will go on till all the clusters are combined into a single cluster.

Remark: In case of large number of cases, icicle plot can be developed by showing cases as rows. For this, specify Horizontal in the Cluster Plots dialog box in SPSS.

The Dendrogram: Plotting Cluster Distances

Figure 10.22 shows the dendrogram which is used to show the plotting of cluster distances. It provides a visual representation of the distance at which clusters are combined. We read the dendrogram from left to right. A vertical line represents the joined clusters. The position of the line on the scale shows the distance at which clusters are joined. The computed distances are rescaled in the range of 1–25, and, therefore, actual distances cannot be seen here; however, the ratio of the rescaled distances within the dendrogram is the same as the ratio of the original distances.

The first vertical line, corresponding to the smallest rescaled distance, is for the case 9 and case 13. The next vertical line is at the next smallest distance for the cluster 9 and case 2. It can be seen from Table 10.8 that the lowest coefficient is

Table 10.9 Initial cluster centers

Variables	Cluster		
	1	2	3
1. The FM station should provide more old Hindi songs	5.00	2.00	4.00
2. FM stations must help an individual in solving their personal problems	3.00	5.00	2.00
3. The presentation style of RJs helps popularizing an FM station	2.00	4.00	3.00
4. An FM station should provide some kind of prizes/incentives to its listeners	4.00	2.00	4.00
5. The station must telecast latest songs more often	5.00	5.00	2.00
6. The FM stations must contain more entertaining programs	2.00	3.00	5.00
7. Popularity of RJs depends upon their humor and ability to make program interesting	4.00	3.00	2.00
8. FM station should provide more opportunity to listeners to talk to celebrities	4.00	5.00	1.00
9. RJs' voice must be clear and melodious	3.00	3.00	5.00
10. FM channels should play 24 × 7	5.00	2.00	2.00
11. FM stations should give information for other sports along with cricket	2.00	5.00	1.00
12. FM stations should provide information regarding educational/professional courses available in the city	5.00	3.00	3.00
13. FM stations should provide information regarding different shopping offers available in the city	2.00	3.00	5.00
14. RJs should speak in an understandable language, preferably in local language	5.00	2.00	3.00

Table 10.10 Iteration history[a]

Iteration	Change in cluster centers		
	1	2	3
1	3.375	1.753	2.953
2	.000	.480	.566
3	.000	.000	.000

[a]Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 7.483

3.000, which is for cases 9 and 13. The next smallest distance is shown by the coefficient as 7.333 which is for cluster 9 and case 2. Thus, what you see in this plot is what you already know from the agglomeration schedule.

Remark: While reading the dendrogram, one should try to determine at what stage the distances between clusters that are combined is large. You look for large distances between sequential vertical lines. In this case, large distance between the vertical lines suggests a three-cluster solution.

Stage 2 With the help of hierarchical cluster analysis, the number of cluster was decided to be three. After this, *K*-means cluster analysis was applied to get the final solution of the cluster means. The SPSS generated the outputs in the form of Tables 10.9, 10.10, 10.11, 10.12, 10.13, and 10.14. We shall now explain these outputs and discuss the cluster characteristics.

Table 10.11 Final cluster centers

Statements	Cluster 1	2	3
1. The FM station should provide more old Hindi songs	4.83[a]	2.13	3.33
2. FM stations must help an individual in solving their personal problems	2.50	4.50[a]	2.67
3. The presentation style of RJs helps popularizing an FM station	2.33	4.25[a]	2.83
4. An FM station should provide some kind of prizes/incentives to its listeners	4.50[a]	2.50	2.50
5. The station must telecast latest songs more often	2.50	4.50[a]	2.50
6. The FM stations must contain more entertaining programs	2.83	2.75	4.33[a]
7. Popularity of RJs depends upon their humor and ability to make program interesting	4.83[a]	2.88	2.83
8. FM station should provide more opportunity to listeners to talk to celebrities	2.83	4.50[a]	2.67
9. RJs' voice must be clear and melodious	2.33	2.50	4.33[a]
10. FM channels should play 24 × 7	4.50[a]	2.88	2.83
11. FM stations should give information for other sports along with cricket	2.33	4.25[a]	2.50
12. FM stations should provide information regarding educational/ professional courses available in the city	4.50[a]	3.13	2.67
13. FM stations should provide information regarding different shopping offers available in the city	2.50	2.88	4.50[a]
14. RJs should speak in an understandable language, preferably in local language	4.33[a]	2.38	2.50

[a]Shows strong agreement toward response

Initial Cluster Centers

The first step in *K*-means clustering was to find the *K*-centers. This is done iteratively. Here, the value of *K* is three because three clusters were decided on the basis of agglomerative schedule. We start with an initial set of centers and keep modifying till the changes between two iterations are small enough. Although one can also guess these centers which can be used as initial starting points, it is advisable to let SPSS find *K* cases that are well separated and use these values as initial cluster centers. In our example, Table 10.9 shows the initial centers.

Once the initial cluster centers are selected by the SPSS, each case is assigned to the nearest cluster, depending upon its distance from the cluster centers. After assigning all the cases to these clusters, the cluster centers are once again recomputed on the basis of its member cases. Again, all the cases are assigned by using the recomputed cluster centers. This process keeps on going till no cluster center changes appreciably. Since the number of iteration is taken as 10 by default in SPSS (see Fig. 10.18), this process of assigning cases and recomputing cluster centers will keep repeating to a maximum of ten times. In this example, you can see from Table 10.10 that the three iterations were sufficient.

Table 10.12 ANOVA table

	Cluster		Error			
	Mean square	df	Mean square	df	*F*	Sig. *p*-value
1. The FM station should provide more old Hindi songs	12.579	2	.885	17	14.217	.000
2. FM stations must help an individual in solving their personal problems	8.858	2	.637	17	13.901	.000
3. The presentation style of RJs helps popularizing an FM station	7.042	2	.333	17	21.125	.000
4. An FM station should provide some kind of prizes/ incentives to its listeners	8.400	2	1.000	17	8.400	.003
5. The station must telecast latest songs more often	9.600	2	1.118	17	8.589	.003
6. The FM stations must contain more entertaining programs	5.042	2	.686	17	7.346	.005
7. Popularity of RJs depends upon their humor and ability to make program interesting	8.204	2	.620	17	13.230	.000
8. FM station should provide more opportunity to listeners to talk to celebrities	7.392	2	.716	17	10.328	.001
9. RJs' voice must be clear and melodious	7.667	2	.745	17	10.289	.001
10. FM channels should play 24 × 7	5.671	2	.777	17	7.299	.005
11. FM stations should give information for other sports along with cricket	8.108	2	.843	17	9.617	.002
12. FM stations should provide information regarding educational/professional courses available in the city	5.546	2	.571	17	9.711	.002
13. FM stations should provide information regarding different shopping offers available in the city	6.938	2	.699	17	9.932	.001
14. RJs should speak in an understandable language, preferably in local language	7.646	2	.512	17	14.926	.000

The *F*-tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal

Final Cluster Centers

Table 10.11 shows the final cluster centers after iteration stops, and cases are reassigned to the clusters. Using these final cluster centers, cluster characteristic are described.

Each question in this example is responded on a 1–5 scoring scale, where 5 stands for total agreement and 1 stands for total disagreement. Thus, if any score shown in Table 10.11 is more than 2.5, it indicates the agreement toward the statement, and if it is less than 2.5, it reflects disagreement. Thus, owing to these

Table 10.13 Cluster membership

Case number	Cluster	Distance
1	1	2.953
2	2	2.378
3	3	2.887
4	1	2.461
5	1	3.424
6	3	3.367
7	3	2.828
8	1	2.779
9	2	2.325
10	2	2.580
11	2	3.828
12	2	3.226
13	2	1.705
14	2	3.487
15	2	2.215
16	1	3.375
17	1	2.838
18	3	3.162
19	3	2.708
20	3	3.317

Table 10.14 Number of cases in each cluster

Cluster	1	6.000
	2	8.000
	3	6.000
Valid		20.000
Missing		.000

criteria, the characteristics of these three clusters of cases were as follows (refer to the question details in Example 10.1):

Cluster 1

FM listeners belonging to this cluster were of the strong opinion that channels should provide more old Hindi songs (Q.1) and provide some incentives to the listeners (Q.4). They strongly feel that the humor and ability to deliver interesting programs make RJs more popular (Q.7). The channel should play 24 × 7 (Q.10) and must air information regarding educational opportunity available in the city (Q.12), and the RJ must speak in local dialect (Q.14).

Further, listeners belonging to this cluster feel that FM channels should air more entertaining programs (Q.6) and should provide more opportunity to listeners to talk to the celebrities (Q.8).

Cluster 2

Listeners belonging to this cluster strongly felt that FM channels must provide solutions to personal problems (Q.2), RJs presentation skill to be important for the channels (Q.3), channels to provide more often the latest songs (Q.5), channels to arrange more dialogues between celebrities and their audience (Q.8), and should air information about sports other than cricket also (Q.11).

Further, listeners to this cluster also felt that FM channels should air more entertaining programs (Q.6). Humor and ability to deliver interesting programs make RJs more popular (Q.7). The channels must play 24 × 7 (Q.10) and should provide information regarding educational opportunity (Q.12) and shopping offers (Q.13) available in the city.

Cluster 3

Listeners in this cluster were strongly of the view that the FM channels must contain more entertaining programs (Q.6), RJs voice must be very clear and melodious (Q.9), and channels should provide information regarding shopping offers available in the city (Q.13).

Further, listeners in this cluster were also of the view that channels should air more old Hindi songs (Q.1), provide solution to the personal problems (Q.2), and believe RJs to be the key factor in popularizing the FM channels (Q.3). They were of the view that the humorous RJs make programs more interesting (Q.7). Channels should provide more opportunity to listeners to talk to the celebrities (Q.8), they should operate 24 × 7 (Q.10) and, at the same time, must air the information regarding educational opportunities available in the city (Q.12).

ANOVA: To Know Differences Between Clusters

Table 10.12 shows ANOVA for the data on all the 14 variables. The F-ratios computed in the table describe the differences between the clusters. F-ratio is significant at 5% level if the significance level (p-value) associated with it is less than .05. Thus, it can be seen in Table 10.12 that F-ratios for all the variables are significant at 5% level as their corresponding p-values are less than .05.

Remark

1. There is a divided opinion on the issue of using ANOVA analysis for comparing the clusters on each of the parameters. The footnote in Table 10.12 warns that the observed significance levels should not be interpreted in the usual fashion because the clusters have been selected to maximize the differences between clusters.

2. It is up to the researcher to decide about using ANOVA for determining the significance of variables. If ANOVA is used, then the interpretation of clusters should be made on the basis of those variables which are significantly different across clusters at any predefined level of significance.

Cluster Membership

Table 10.13 shows the cluster membership of the cases. You can see that six cases belong to cluster 1, eight cases to cluster 2, and six cases to cluster 3.

Table 10.14 is a summary of Table 10.13. You do not like to see clusters with very few cases unless they are really different from the remaining cases.

Exercise

Short Answer Questions

Q.1. Discuss a research situation where cluster analysis can be applied.
Q.2. Write the steps involved in cluster analysis.
Q.3. What is squared Euclidean distance? What impact it will have if the variables are measured on different units? Suggest the procedure in that situation.
Q.4. When should you use Chebyshev distance and Mahalanobis distance? How these distances are computed?
Q.5. How hierarchical clustering is different than *K*-means clustering?
Q.6. What is the difference between single linkage and average linkage method?
Q.7. What do you mean by Ward's minimum variance method?
Q.8. What is the difference between agglomerative and divisive clustering? Can both these clustering be shown in a single graph? If yes, how?
Q.9. In what situation two-stage clustering is done? Explain the steps in brief.
Q.10. Why hierarchical clustering is known as explorative technique? Explain briefly the advantage of using this method.
Q.11. In cluster analysis, when do we need to standardize the variable and why?
Q.12. What do you mean by icicle plot and what it conveys? Show it by sketch.
Q.13. What is the purpose of proximity matrix? Develop a proximity matrix by using any set of data on three cases measured on four variables.
Q.14. Discuss the assumptions used in cluster analysis.
Q.15. How the properties of clusters are explained?
Q.16. Would you agree to use ANOVA in cluster analysis? If yes, how clusters should be explained, and, if not, why?
Q.17. In using SPSS for *K*-means cluster analysis, what output would be generated if the option "Cluster interaction for each case" is chosen?

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers. Tick mark the one that you consider the closest to the correct answer.

Answer Q.1 to Q.4 on the basis of the following information. In a cluster analysis, if the data on two cases are as follows:

Case 1	14	8	10
Case 2	10	11	12

1. The squared Euclidean distance shall be

   (a) 27
   (b) 29
   (c) 28
   (d) $\sqrt{29}$

2. The Manhattan distance shall be

   (a) 81
   (b) 9
   (c) 3
   (d) $\sqrt{9}$

3. The Chebyshev distance shall be

   (a) 4
   (b) 3
   (c) 2
   (d) 9

4. The Euclidean distance shall be

   (a) 29
   (b) 30
   (c) $\sqrt{29}$
   (d) 28

5. Cluster analysis is a(n)

   (a) Explorative analysis
   (b) Descriptive analysis
   (c) Deductive analysis
   (d) Predictive analysis

6. When cluster is formed in agglomerative clustering by joining case 3 with case 7, then the resultant cluster would be known as

   (a) Cluster 7
   (b) Cluster 3
   (c) Cluster 3,7
   (d) Cluster 7,3

7. In Ward's method, any two clusters are joined to form a new cluster if

   (a) The variation within each cluster is same
   (b) The variation within one cluster is minimum than other
   (c) The variation within the two clusters is least
   (d) The variation between both the clusters is maximum

8. In complete linkage method, the clusters are formed on the basis of

   (a) Minimum distance between the closest members of the two clusters
   (b) Minimum average distance between all pairs of objects (in each pair, one member must be from a different cluster)
   (c) Minimum square Euclidean distances between any two clusters
   (d) Minimum distance between the farthest members of the two clusters

9. The number of clusters is decided on the basis of fusion coefficients in agglomerative schedule. In doing so, if distance matrix is considered, then

   (a) We look for the largest difference in the coefficients
   (b) We look for the smallest difference in the coefficients
   (c) We look for the equality of the coefficients
   (d) It is up to the researcher to decide any criteria

10. In cluster analysis, the characteristics of the clusters are decided on the basis of

    (a) Initial cluster solution
    (b) Final cluster solution
    (c) Cluster membership
    (d) ANOVA table

Assignments

1. The data on five nutritional contents of different food articles are shown in Table-A. Identify the suitable clusters of food articles based on these five nutritional contents. Use centroid clustering method and squared Euclidean distances to find the clusters. Apply hierarchical clustering and then *K*-means clustering method to find the final solution for clusters. Explain your findings and discuss the characteristics of different clusters.
2. Ratings were obtained on different brands of car for their six parameters shown in the Table-B. These cars are in specific prize range. The rating 1 indicates complete agreement and 5 indicates complete disagreement. Apply cluster analysis to discuss the characteristics of identified clusters of cars. Use Ward's method of clustering and squared Euclidean distance measure for cluster formation. Use the label cases option in SPSS.

Table-A Nutritional components of different food articles

Food article	Carbohydrates	Protein	Fat	Iron	Vitamin
1	354	20	28	10	2.4
2	89	13	3	38	1.7
3	375	20	33	8	2.6
4	192	23	11	17	3.7
5	116	21	13	12	1.8
6	169	24	8	12	1.5
7	160	18	10	115	2.5
8	320	24	16	9	2.9
9	220	8	31	12	2.5
10	158	25	6	11	5.9
11	202	19	15	7	2.5
12	263	21	21	9	2.8
13	280	21	29	10	2.8
14	72	12	3	83	6
15	46	8	6	74	5.4
16	415	14	40	7	2
17	132	18	4	14	3.5
18	204	20	12	6	1
19	125	11	40	12	2.3
20	342	21	27	8	2.5
21	189	22	10	9	2.7
22	136	23	5	22	2.8

(Hint: Transform your data into *Z*-scores by using the commands in SPSS)

Table-B Ratings on different cars on their characteristics

S.N.	Car	1	2	3	4	5	6
1	Logan	4	2	2	4	4	4
2	Renault Logan Edge	4	3	3	4	3	3
3	Mahindra-Renault Logan Edge	3	2	4	2	4	3
4	Mahindra Verito	4	4	2	3	3	3
5	Swift Dzire	3	3	3	2	2	3
6	Maruti Swift Dzire	4	2	2	3	3	4
7	Chevrolet Beat	4	3	1	4	2	5
8	Tata Venture	5	2	2	3	4	2
9	Chevrolet Aveo	4	3	3	2	3	1
10	Tata Sumo Spacio	2	4	2	3	3	2
11	Skoda New Fabia	3	5	3	5	4	4
12	Hyundai i10	4	4	3	4	3	3
13	Tata Indigo e-CS	3	3	4	3	3	2
14	Maruti Suzuki Swift	4	4	3	2	4	4
15	Maruti Suzuki A-Star	2	5	2	4	2	1
16	Maruti Suzuki Ritz Genus	3	4	3	5	5	2
17	Premier Rio	1	3	4	2	3	2
18	Nissan Micra	2	2	4	3	3	4
19	Volkswagen Polo	3	3	4	5	5	5
20	Skoda Fabia	4	1	5	4	4	5
21	Mahindra-Renault Logan	3	3	4	3	2	4
22	Tata Sumo Victa	2	2	5	3	3	3
23	Tata Sumo Grande	3	4	4	2	2	2
24	Tata Indigo Marina	4	2	5	3	3	1

Parameters of the Car

1. The leg space in the car is comfortable.
2. Car space is big enough to keep my luggage during outing.
3. The car is giving the same mileage as mentioned in the brochure.
4. Driving is very comfortable.
5. Security feature of the car is good.
6. Accessories provided in the car are of good quality.

Besides explaining the characteristics of clusters, also answer the following:

(a) What are the minimum and maximum distances between the cases?
(b) How many clusters you would like to identify and what is the maximum distance between the fusion coefficients?
(c) What criteria you would adopt to discuss the properties of the clusters?
(d) Explain cluster characteristics on the basis of ANOVA and see if it is different than what you have explained earlier.
(e) How many cases/cars are in each cluster?

Answers to Multiple-Choice Questions

Q.1 b	Q.2 b	Q.3 a	Q.4 c
Q.5 a	Q.6 b	Q.7 c	Q.8 d
Q.9 a	Q.10 b		