Block 18: Cluster analysis

(Activity solutions can be found at the end of the document.)

Like factor analysis, **cluster analysis** also examines an entire set of interdependent relationships. The primary objective of cluster analysis is to *classify objects into relatively homogeneous groups* based on the variables. In essence, cluster analysis is the obverse of factor analysis - *reducing the number of objects* rather than variables. Clustering is very useful for defining *target markets*.

Learning Objectives

- describe the basic concept and scope of cluster analysis and its importance in market research
- discuss the statistics associated with cluster analysis
- explain the procedure for conducting cluster analysis, including formulating the problem, selecting a distance measure, selecting a clustering procedure, deciding on the number of clusters, interpreting clusters and profiling clusters
- describe the purpose and methods for evaluating the quality of clsutering results and assessing reliability and validity.

Readings

Malhotra, N.K., D. Nunan and D.F. Birks. *Marketing Research: An Applied Approach.* (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 25.

18.1 Cluster analysis

For each section of *Cluster analysis*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Overview of cluster analysis

Cluster analysis is a class of techniques used to classify objects or cases into relatively homogeneous groups called **clusters**. Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters. Cluster analysis is also called *classification analysis* or *numerical taxonomy*.

Both cluster analysis and discriminant analysis are concerned with *classification*. However, discriminant analysis requires prior knowledge of the cluster or group membership for each object or case included, to develop the classification rule. In contrast, in cluster analysis there is *no a priori information* about the group or cluster membership for any of the objects. Groups or *clusters are suggested by the data*, not defined *a priori*.

Figure 25.1 of the textbook shows an ideal clustering solution, while Figure 25.2 of the textbook shows a (more realistic) practical clustering solution.

Activity 18.1

Discuss the similarity and difference between cluster analysis and discriminant analysis.

Activity 18.2

What is a ‘cluster’?

Uses of cluster analysis

Segmenting the market: Recognising customers' differences is the key to successful marketing, which can lead to a closer matching between products and customer needs. Consumers may be clustered on the basis of benefits sought from the purchase of a product (*benefit segmentation*).

Understanding buyer behaviour: For example, what kind of strategies do car buyers use for buying a car?

Identifying new product opportunities: Clustering brands and products so that competitive sets within the market can be determined.

Selecting test markets: Grouping cities into homogeneous clusters in order to test various marketing strategies.

Reducing data: Achieve simplicity through reducing the original dimensionality of the data.

However, note that cluster analysis is a *distribution-free method*.

Activity 18.3

What are some of the uses of cluster analysis in marketing?

Statistics associated with cluster analysis

Agglomeration schedule - An agglomeration schedule gives information on the objects or cases being combined at each stage of a hierarchical clustering process.

Cluster centroid - The cluster centroid is the mean values of the variables for all the cases or objects in a particular cluster.

Cluster centres - The cluster centres are the initial starting points in non-hierarchical clustering. Clusters are built around these centres, or *seeds*.

Cluster membership - Cluster membership indicates the cluster to which each object or case belongs.

Dendrogram - A dendrogram, or tree graph, is a graphical device for displaying clustering results. Vertical lines represent clusters which are joined together. The position of the line on the horizontal scale indicates the distance at which clusters are joined. The dendrogram is read from left to right.

Distances between cluster centres - These distances indicate how separated the individual pairs of clusters are. Clusters which are widely separated are distinct and, therefore, desirable.

Icicle diagram - An icicle diagram is a graphical display of clustering results, so called because it resembles a row of icicles hanging from the eaves of a house. The columns correspond to the objects being clustered, and the rows correspond to the number of clusters. An icicle diagram is read from top to bottom.

Similarity/distance coefficient matrix - A similarity/distance coefficient matrix is a lower triangular matrix containing pairwise distances between objects or cases.

For four objects, we have a 4×4 **distance matrix**:

$$\begin{pmatrix} - & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & - & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & - & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & - \end{pmatrix}$$

where δ_{ij} is the distance between object *i* and object *j*. Clearly, δ_{12}= δ_{21}, δ_{13}= δ_{31} etc. The diagonal may be left blank.

Activity 18.4

Briefly define the following terms: dendrogram, icicle plot, agglomeration schedule and cluster membership.

Conducting cluster analysis

The process to conduct cluster analysis is as follows:

Formulate the problem

⇓⇓

Select a distance measure

⇓⇓

Select a clustering procedure

⇓⇓

Decide on the number of clusters

⇓⇓

Interpret and profile clusters

⇓⇓

Assess the reliability and validity

Conducting cluster analysis

Perhaps the most important part of formulating the clustering problem is *selecting the variables* on which the clustering is based. Inclusion of even one or two irrelevant variables may distort an otherwise useful clustering solution. Basically, the set of variables selected should *describe the similarity between objects in terms which are relevant to the market research problem*. The variables should be selected based on past research, theory or a consideration of the hypotheses being tested. In exploratory research, the researcher should activity judgement and intuition.

The most commonly-used measure of similarity is the **Euclidean distance**, or its square. The Euclidean distance is the square root of the sum of the squared differences in values for each variable. Other distance measures are also available. The **city-block** or **Manhattan distance** between two objects is the sum of the absolute differences in values for each variable. The **Chebychev distance**between two objects is the maximum absolute difference in values for any variable.

The Euclidean distance between objects ii and jj is:

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

For p=2, the Euclidean distance corresponds to the 'straight line' distance between the two points $(x_{i1}x_{i2})$ and (x_{j1}, x_{j2}).

A *weight*, w_k could be assigned to variable *k* if it was believed that more importance should be attached to some variables over others giving:

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

The city-block measure is:

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

Compared to Euclidean distance, this gives less relative weight to large differences.

The Chebychev distance is:

$$\delta_{ij} = \max_{\forall k} \left| x_{ik=x_{jk})} \right|$$

If the variables are measured in vastly different units, the *clustering solution will be influenced by the units of measurement*. In these cases, before clustering participants, we must *standardise the data* by rescaling each variable to have a mean of zero and a standard deviation of one. It is also desirable to *eliminate outliers* (cases with atypical values). The use of different distance measures may lead to different clustering results. Hence it is advisable to *use different measures and compare the results*.

Activity 18.5

What is the most commonly-used measure of similarity in cluster analysis?

Clustering procedures

Figure 25.4 of the textbook provides a classification of clustering procedures.

Hierarchical clustering is characterised by the development of a hierarchy or tree-like structure. Hierarchical methods can be agglomerative or divisive.

Agglomerative clustering starts with each object in a separate cluster. Clusters are formed by grouping objects into larger and larger clusters. This process is continued until all objects are members of a single cluster.

Divisive clustering starts with all the objects grouped in a single cluster. Clusters are divided or split until each object is in a separate cluster.

Agglomerative methods are commonly used in market research. They consist of linkage methods, error sums of squares or variance methods and centroid methods.

The **single linkage method** is based on the minimum distance or the nearest neighbour rule. At every stage, the distance between two clusters is the distance between their two closest points.

The **complete linkage method** is similar to single linkage, except that it is based on the maximum distance or the farthest neighbour approach. In complete linkage, the distance between two clusters is calculated as the distance between their two farthest points.

The **average linkage method** works similarly. However, in this method, the distance between two clusters is defined as the average of the distances between all pairs of objects, where one member of the pair is from each of the clusters.

Figure 25.5 of the textbook shows the linkage methods of clustering.

A **variance method** attempts to generate clusters to *minimise the within-cluster variance*. A commonly-used variance method is **Ward's procedure**. For each cluster, the means for all the variables are computed. Next, for each object, the squared Euclidean distance to the cluster means is calculated. These distances are summed for all the objects. At each stage, the two clusters with the smallest increase in the overall sum of squares within cluster distances are combined.

In the **centroid method**, the distance between two clusters is the distance between their centroids (means for all the variables). Every time objects are grouped, a new centroid is computed.

Of the hierarchical methods, average linkage and Ward's methods have been shown to perform better than the other procedures.

Figure 25.6 of the textbook shows other agglomerative clustering methods.

The **non-hierarchical clustering methods** are frequently referred to as kk-means clustering. These methods include sequential threshold, parallel threshold and optimising partitioning.

In the **sequential threshold method**, a cluster centre is selected and all objects within a pre-specified threshold value from the centre are grouped together. Next a new cluster centre or seed is selected, and the process is repeated for the unclustered points. Once an object is clustered with a seed, it is no longer considered for clustering with subsequent seeds.

The **parallel threshold method** operates similarly, except that several cluster centres are selected simultaneously and objects within the threshold level are grouped with the nearest centre.

The **optimising partitioning method** differs from the two threshold procedures in that objects can later be reassigned to clusters to optimise an overall criterion, such as average within-cluster distance for a given number of clusters.

It has been suggested that the hierarchical and non-hierarchical methods be used in tandem. First, an *initial clustering solution is obtained using a hierarchical procedure*, such as average linkage or Ward's. The number of clusters and cluster centroids so obtained are used as inputs to the optimising partitioning method.

The choice of a clustering method and the choice of a distance measure are interrelated. For example, squared Euclidean distances should be used with Ward's procedure and centroid methods. Several non-hierarchical procedures also use squared Euclidean distances.

Activity 18.6

Present a classification of clustering procedures.

Activity 18.7

On what basis may a researcher decide which variables should be selected to formulate a clustering problem?

Activity 18.8

Why is the average linkage method usually preferred to single linkage and complete linkage?

Activity 18.9

What are the two major disadvantages of non-hierarchical clustering procedures?

Deciding clusters, interpretation, profiling, reliability and validity

Theoretical, conceptual or practical considerations may suggest a certain number of clusters. In hierarchical clustering, the distances at which clusters are combined can be used as criteria. This information can be obtained from the agglomeration schedule or from the dendrogram. In non-hierarchical clustering, the ratio of total within-group variance to between-group variance can be plotted against the number of clusters. The point at which an elbow or a sharp bend occurs indicates an appropriate number of clusters. The *relative sizes* of the clusters should be meaningful.

Interpreting and profiling clusters involves *examining the cluster centroids*. The centroids enable us to describe each cluster by assigning it a name or label. It is often helpful to *profile the clusters* in terms of variables which were not used for clustering. These may include demographic, psychographic, product usage, media usage or other variables.

Reliability and validity can be assessed in the following ways.

- Perform cluster analysis on the same data using different distance measures. Compare the results across measures to *determine the stability of the solutions*.
- Use different methods of clustering and compare the results.
- Split the data randomly into halves. Perform clustering separately on each half. Compare cluster centroids across the two subsamples.
- Delete variables randomly. Perform clustering based on the reduced set of variables. Compare the results with those obtained by clustering based on the entire set of variables.
- In non-hierarchical clustering, the solution may depend on the order of cases in the dataset. Make multiple runs using a different order of cases until the solution stabilises.

It is also possible to cluster variables. In this instance, the units used for analysis are the variables, and the distance measures are computed for all pairs of variables. Hierarchical clustering of variables can aid in the identification of unique variables, or variables which make a unique contribution to the data. Clustering can also be used to *reduce the number of variables*. Associated with each cluster is a *linear combination of the variables in the cluster*, called the *cluster component*. A large set of variables can often be replaced by the set of cluster components with little loss of information. However, a given number of cluster components does not generally explain as much variance as the same number of principal components.

Activity 18.10

What guidelines are available for deciding on the number of clusters?

Activity 18.11

What is involved in the interpretation of clusters?

Activity 18.12

What role may qualitative methods play in the interpretation of clusters?

Activity 18.13

What are some of the additional variables used for profiling the clusters?

Activity 18.14
Describe some procedures available for assessing the quality of clustering solutions.

Activity 18.15
How is cluster analysis used to group variables?

Cluster analysis example: Shopping attitudes
Consider a clustering of consumers based on attitudes toward shopping. Suppose previous research has identified six attitudinal variables as being the most relevant to shopping attitudes. 20 participants were asked to express their degree of agreement with the following statements on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

- V_1 = Shopping is fun.
- V_2 = Shopping is bad for your budget.
- V_3 = I combine shopping with eating out.
- V_4 = I try to get the best buys while shopping.
- V_5 = I don't care about shopping.
- V_6 = You can save a lot of money by comparing prices.

Table 25.1 of the textbook provides the attitudinal data for this clustering example. The data can be downloaded from the file *Shopping.sav* or can be seen in the table below.

Number	V1	V2	V3	V4	V5	V6
1	6	4	7	3	2	3
2	2	3	1	4	5	4
3	7	2	6	4	1	3
4	4	6	4	5	3	6
5	1	3	2	2	6	4
6	6	4	6	3	3	4
7	5	3	6	3	3	4
8	7	3	7	4	1	4
9	2	4	3	3	6	3
10	3	5	3	6	4	6
11	1	3	2	3	5	3
12	5	4	5	4	2	4
13	2	2	1	5	4	4
14	4	6	4	6	4	7
15	6	5	4	2	1	4
16	3	5	4	6	4	7
17	4	4	7	2	2	5
18	3	7	2	6	4	3
19	4	6	3	7	2	7
20	2	3	2	4	7	2

Table 25.2 of the textbook shows the results of hierarchical clustering using squared Euclidean distance and Ward's procedure. The output includes the agglomeration schedule that reports which clusters were combined at which stage of the agglomeration, as well as cluster membership for two, three and four cluster solutions.

Figure 25.7 of the textbook shows the vertical icicle plot for this cluster analysis, which details the cases belonging to each cluster for different numbers of clusters.

Figure 25.8 of the textbook provides the dendrogram which suggests a three-cluster solution may be appropriate.

Table 25.3 of the textbook shows the cluster centroids (i.e. the means of variables) for the three-cluster solution. Cluster centroids are useful for profiling clusters as the mean values indicate a representative member of each cluster.

Cluster analysis example: Supermarket customers

Suppose we examine five customers of a supermarket: Adam, Brian, Carmen, Donna and Eve (denoted A, B, C, D and E, respectively). The entries below represent how far apart individuals are in regard to their potential buying patterns.

	A	**B**	**C**	**D**	**E**
A	-				
B	3	-			
C	8	7	-		
D	11	9	6	-	
E	10	9	7	5	-

Using the nearest neighbour (single linkage) method, first we look for the closest pair of individuals, i.e. Adam and Brian. We construct a new distance table appropriate for the four clusters existing at the end of the first stage - the distance between two clusters is defined as the distance between their nearest members.

	(A, B)	**C**	**D**	**E**
(A, B)	-			
C	7	-		
D	9	6	-	
E	9	7	5	-

We repeat the procedure until all objects are in one cluster. More specifically, the next most similar pair of objects is Donna and Eve.

	(A, B)	**C**	**(D, E)**
(A, B)	-		
C	7	-	
(D, E)	9	6	-

The agglomeration schedule is:

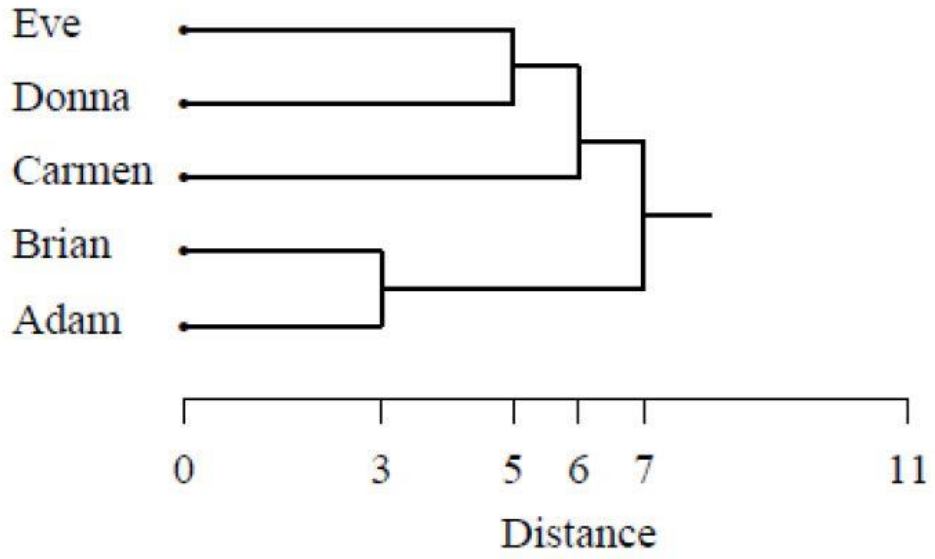
Stage	**Number of clusters**	**Clusters**	**Distance level**
Initial	5	(A) (B) (C) (D) (E)	0
1	4	(A, B) (C) (D) (E)	3
2	3	(A, B) (C) (D, E)	5
3	2	(A, B) (C, D, E)	6
4	1	(A, B, C, D, E)	7

Using the farthest neighbour (complete linkage) method, first we look for the most remote pair of individuals. The agglomeration schedule is:

Stage	**Number of clusters**	**Clusters**	**Distance level**
Initial	5	(A) (B) (C) (D) (E)	0
1	4	(A, B) (C) (D) (E)	3
2	3	(A, B) (C) (D, E)	5
3	2	(A, B) (C, D, E)	7
4	1	(A, B, C, D, E)	11

The sets of clusters produced by the farthest neighbour method coincide with those from the nearest neighbour method, although the distance levels differ at which the clusters merge. Generally, the nearest and farthest neighbour methods give different results, sometimes very different, especially when there are many objects or individuals to be clustered. Both methods only depend on the ordinal properties of the distances.

The dendrogram for the nearest neighbour (single linkage) clustering is:



The dendrogram for the farthest neighbour (complete linkage) clustering is:

Discussion forum, activities and discussion points

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Activities on the block's topics

1. Replicate the cluster analysis results from the lecture example using the data file *Shopping.sav*. (An Excel version of the dataset is *Shopping.xlsx*.)

Video walkthrough of activity 1.

2. In a survey pretest, data were obtained from 20 participants on preference for boots on a seven-point scale (1 = not preferred, 7 = greatly preferred) (V1V1). The participants also provided their evaluations of the boots on comfort (V2V2), style (V3V3) and durability (V4V4), also on seven-point scales (1 = poor, 7 = excellent). The resulting data are given in the file *Boots.sav*. (An Excel version of the dataset is *Boots.xlsx*.)

   Conduct cluster analysis on the boots data. Consider only the following variables: evaluations of the boots on comfort (V2V2), style (V3V3) and durability (V4V4).

   Video walkthrough of activity 2.

3. In a survey pre-test, data were obtained from 45 participants on Benetton clothes. These data are given in the file *Benetton.sav*, which gives the usage, gender, awareness, attitude, preference, intention and loyalty toward Benetton of a sample of Benetton users. Usage was coded as 1, 2 or 3, representing light, medium or heavy users, respectively. Gender was coded as 1 for females and 2 for males. Awareness, attitude, preference, intention and loyalty were measured on a 7-point Likert-type scale (1 = Very unfavourable, 7 = Very favourable). Note that five participants have missing values which are denoted by 9. (An Excel version of the dataset is *Benetton.xlsx*.)

   Conduct cluster analysis on the Benetton data. Consider only the following variables: awareness, attitude, preference, intention and loyalty toward Benetton.

   Video walkthrough of activity 3.

Discussion points

1. 'The consequences of inappropriate validation of cluster analysis solutions can be disastrous.'
2. 'User-friendly statistical packages can create cluster solutions in situations where naturally-occurring clusters do not exist.'

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- Describe the basic concept and scope of cluster analysis and its importance in market research
- Discuss the statistics associated with cluster analysis
- Explain the procedure for conducting cluster analysis, including formulating the problem, selecting a distance measure, selecting a clustering procedure, deciding on the number of clusters, interpreting clusters and profiling clusters
- Describe the purpose and methods for evaluating the quality of clustering results and assessing reliability and validity.

Block 18: Cluster analysis

Solution to Activity 18.1

Both cluster analysis and discriminant analysis are concerned with the classification of objects, cases or variables into relatively homogeneous groups. However, in cluster analysis the groups or clusters are suggested by the data, whereas in discriminant analysis they are defined *a priori*.

Solution to Activity 18.2

If a dataset is analysed with a cluster analysis programme it will produce clustering solutions. In other words, the analysis will show clusters whether there are 'naturally-occurring' in the dataset or not. A naturally-occurring cluster seen in two dimensions may be circular, crescent-shaped, mushroom-shaped, cigar-shaped or any other shape. In three dimensions it may be circular, cylindrical, conical or any other shape. Beyond three dimensions the cluster may not be observed, only two- or three-dimensional 'slices' through the cluster can be observed. The analysis package 'suggests' cluster solutions - the researcher has to go through steps to validate each 'suggestion' and demonstrate that it is a naturally-occurring cluster.

In essence, the cluster has characteristics which make cluster members 'alike' or homogeneous and 'different' or distinct from other members.

Solution to Activity 18.3

The technique has several applications in marketing. Some of them are as follows.

- *Segmenting the market*. Using this technique, consumers are clustered on the basis of benefits sought from the purchase of a particular product.
- *Understanding buyer behaviour*. Cluster analysis is used to classify homogeneous groups of buyers to facilitate the study of buyer behaviour in each group.
- *Identifying new product opportunities*. By using this technique, a firm can cluster the brands/products and examine its current offerings compared to those of its competitors to identify potential new product opportunities.
- *Selecting test markets*. It can be used to group cities into homogeneous clusters and then select comparable cities to test various marketing strategies.
- *Reducing data*. It may be utilised to develop clusters or subgroups of data which are more manageable than individual observations, to be used in subsequent multivariate analyses.

Solution to Activity 18.4

- *Dendrogram*. Also called a tree graph, a dendrogram is a graphical device for displaying clustering results and is read from left to right. Vertical lines represent clusters which are joined together. Distances at which the clusters are joined are indicated by the position of the line on the horizontal scale.

- *Icicle plot*. A graphical display of clustering results which is read from top to bottom. The columns correspond to the objects being clustered and the rows correspond to the number of clusters.
- *Agglomeration schedule*. This table gives information on the objects or cases being combined at each stage of a hierarchical clustering process.
- *Cluster membership*. This indicates the cluster to which each object or case belongs.

Solution to Activity 18.5

The Euclidean distance or its square is the most commonly-used measure of similarity. This is based on the common approach of measuring similarity in terms of distance between pairs of objects.

Solution to Activity 18.6

The clustering procedures can be divided into two major categories: hierarchical and non-hierarchical. Hierarchical clustering is characterised by the development of a hierarchy or tree-like structures. The non-hierarchical methods are also called kk-means clustering.

Solution to Activity 18.7

Perhaps the most important part of formulating the clustering problem is to select the variables on which the clustering is based. Inclusion of even one or two irrelevant variables may distort an otherwise useful clustering solution. Basically, the set of variables selected should describe the similarity between objects in terms which are relevant to the market research problem. The variables should be selected based on past research, theory or a consideration of the hypotheses being tested. As clustering is an exploratory approach, the researcher should also activity judgement and intuition in the choice of variables.

Solution to Activity 18.8

The average linkage method is most preferable because it uses information on all pairs of distances and not just the maximum and minimum distances. Therefore, it performs better than the single and complete linkage methods.

Solution to Activity 18.9

In non-hierarchical procedures, the number of clusters has to be pre-specified and the selection of cluster centres is arbitrary, which can lead to potential biases.

Solution to Activity 18.10

The number of clusters may be decided by any of the following procedures.

Theoretical, conceptual or practical considerations may dictate the choice of the number of clusters.

In hierarchical clustering, the distances at which clusters are being combined have a large bearing on the choice of the number of clusters.

In non-hierarchical clustering, the ratio of total within-group variance to between-group variance is plotted against the number of clusters. The maximum number of clusters which should be selected is then indicated by the point where a sharp bend occurs.

The number of clusters should be selected so as to make the relative size of the clusters meaningful.

Solution to Activity 18.11

The clusters should be interpreted in terms of cluster centroids which represent the average values of the objects contained in the cluster on each of the variables. This enables the researcher to describe each cluster by assigning it a name or label, which may be obtained from the cluster programme or through discriminant analysis.

Solution to Activity 18.12

A clustering solution may suggest clustering based on members being alike in some respects and different in other respects. The analysis is based on exploring the dataset, looking for common similarities and differences. When a solution is suggested, the researcher must ask whether the common similarities and differences make any sense.

As well as examining the logic of connecting variables which make up a cluster, the researcher should examine the means to classify or name the connections which make a distinct cluster. This process can be achieved through qualitative research techniques. Participants can be asked to respond to suggested cluster structures, to 'play around' with alternative structures and names for structures or clusters. This process can be done on a one-to-one basis or in focus groups.

Solution to Activity 18.13

Other additional variables used for profiling clusters include psychographic, demographic, product usage, media usage and other variables.

Solution to Activity 18.14

The reliability and validity of the clustering solutions may be checked with the following procedures.

- Comparison of cluster analysis results obtained on the same data using different distance measures determines the stability of the solutions.
- Comparison of results obtained with different methods can also serve as a check.
- Often, clustering is done separately on each half of the data split randomly into two halves. Subsequently, the cluster centroids can be compared across the two subsamples.
- Sometimes clustering is done on a reduced set of variables (using a random deletion of variables) and compared with those obtained by clustering based on the entire set of variables.
- In non-hierarchical clustering, the solution may depend on the order of cases in the dataset. Multiple runs should be made using different orders of cases until the solution stabilises.

Solution to Activity 18.15

Cluster analysis is used for grouping variables to identify the homogeneous groups of variables. While clustering the variables, the units used for analysis are the variables, and the distance measures should be computed for all pairs of variables.
As an example, the correlation coefficient (either the absolute value or with the sign) may be used as a measure of similarity (as opposed to distance) between the variables.

Solutions to activitys on the block's topics

1. SPSS output from running cluster analysis on the dataset is as follows.

Case Processing Summary[a]

Cases					
Valid		**Missing**		**Total**	
N	**Percent**	**N**	**Percent**	**N**	**Percent**
20	100.0	0	0.0	20	100.0

a. Ward Linkage

This first box confirms the number of observations used in the cluster analysis and reports any missing values (here there are no missing values).

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	64.000	8.000	31.000	69.000	3.000	5.000	5.000	48.000	48.000	60.000	7.000	65.000	46.000	13.000	48.000	9.000	56.000	56.000	69.000
2	64.000	.000	68.000	31.000	7.000	47.000	39.000	77.000	8.000	18.000	4.000	35.000	3.000	36.000	49.000	28.000	55.000	24.000	44.000	9.000
3	8.000	68.000	.000	43.000	83.000	11.000	11.000	3.000	64.000	56.000	70.000	11.000	61.000	58.000	19.000	58.000	23.000	70.000	60.000	79.000
4	31.000	31.000	43.000	.000	44.000	20.000	22.000	36.000	31.000	5.000	39.000	12.000	34.000	3.000	22.000	5.000	24.000	17.000	7.000	50.000
5	69.000	7.000	83.000	44.000	.000	52.000	42.000	90.000	5.000	33.000	3.000	46.000	16.000	51.000	58.000	41.000	52.000	41.000	69.000	10.000
6	3.000	47.000	11.000	20.000	52.000	.000	2.000	8.000	35.000	33.000	47.000	4.000	50.000	31.000	10.000	33.000	8.000	45.000	43.000	54.000
7	5.000	39.000	11.000	22.000	42.000	2.000	.000	10.000	29.000	31.000	37.000	4.000	40.000	33.000	14.000	31.000	6.000	47.000	45.000	46.000
8	5.000	77.000	3.000	36.000	90.000	8.000	10.000	.000	69.000	53.000	79.000	10.000	72.000	49.000	18.000	51.000	16.000	71.000	53.000	90.000
9	48.000	8.000	64.000	31.000	5.000	35.000	29.000	69.000	.000	24.000	4.000	31.000	17.000	38.000	45.000	32.000	41.000	24.000	56.000	5.000
10	48.000	18.000	56.000	5.000	33.000	33.000	31.000	53.000	24.000	.000	28.000	21.000	19.000	4.000	39.000	2.000	39.000	14.000	8.000	35.000
11	60.000	4.000	70.000	39.000	3.000	47.000	37.000	79.000	4.000	28.000	.000	37.000	9.000	48.000	51.000	38.000	49.000	30.000	60.000	7.000
12	7.000	35.000	11.000	12.000	46.000	4.000	4.000	10.000	31.000	21.000	37.000	.000	34.000	23.000	8.000	23.000	10.000	31.000	27.000	48.000
13	65.000	3.000	61.000	34.000	16.000	50.000	40.000	72.000	17.000	19.000	9.000	34.000	.000	39.000	52.000	29.000	58.000	29.000	41.000	16.000
14	46.000	36.000	58.000	3.000	51.000	31.000	33.000	49.000	38.000	4.000	48.000	23.000	39.000	.000	39.000	2.000	37.000	22.000	6.000	55.000
15	13.000	49.000	19.000	22.000	58.000	10.000	14.000	18.000	45.000	39.000	51.000	8.000	52.000	39.000	.000	43.000	16.000	43.000	41.000	68.000
16	48.000	28.000	58.000	5.000	41.000	33.000	31.000	51.000	32.000	2.000	38.000	23.000	29.000	2.000	43.000	.000	35.000	24.000	8.000	47.000
17	9.000	55.000	23.000	24.000	52.000	8.000	6.000	16.000	41.000	39.000	49.000	10.000	58.000	37.000	16.000	35.000	.000	59.000	49.000	68.000
18	56.000	24.000	70.000	17.000	41.000	45.000	47.000	71.000	24.000	14.000	30.000	31.000	29.000	22.000	43.000	24.000	59.000	.000	24.000	31.000
19	56.000	44.000	60.000	7.000	69.000	43.000	45.000	53.000	56.000	8.000	60.000	27.000	41.000	6.000	41.000	8.000	49.000	24.000	.000	73.000
20	69.000	9.000	79.000	50.000	10.000	54.000	46.000	90.000	5.000	35.000	7.000	48.000	16.000	55.000	68.000	47.000	68.000	31.000	73.000	.000

This is a dissimilarity matrix

The proximity matrix reports how far apart each pair of observations is based on the distance measure specified when running the cluster analysis (here squared Euclidean distance was used). Hierarchical clustering procedures base the agglomeration by combining objects which are in closest proximity to each other.

Clearly, for large numbers of observations, the proximity matrix becomes very large so typically would not be reported. Nevertheless, when reviewing the agglomeration schedule it can be worthwhile to cross-reference this with the proximity matrix.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	1.000	0	0	6
2	6	7	2.000	0	0	7
3	2	13	3.500	0	0	15
4	5	11	5.000	0	0	11
5	3	8	6.500	0	0	16
6	10	14	8.167	0	1	9

7	6	12	10.500	2	0	10
8	9	20	13.000	0	0	11
9	4	10	15.583	0	6	12
10	1	6	18.500	0	7	13
11	5	9	23.000	4	8	15
12	4	19	27.750	9	0	17
13	1	17	33.100	10	0	14
14	1	15	41.333	13	0	16
15	2	5	51.833	3	11	18
16	1	3	64.500	14	5	19
17	4	18	79.667	12	0	18
18	2	4	172.667	15	17	19
19	1	2	328.600	16	18	0

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	1.000	0	0	6
2	6	7	2.000	0	0	7
3	2	13	3.500	0	0	15
4	5	11	5.000	0	0	11
5	3	8	6.500	0	0	16
6	10	14	8.167	0	1	9
7	6	12	10.500	2	0	10
8	9	20	13.000	0	0	11
9	4	10	15.583	0	6	12
10	1	6	18.500	0	7	13
11	5	9	23.000	4	8	15
12	4	19	27.750	9	0	17
13	1	17	33.100	10	0	14
14	1	15	41.333	13	0	16
15	2	5	51.833	3	11	18
16	1	3	64.500	14	5	19
17	4	18	79.667	12	0	18
18	2	4	172.667	15	17	19
19	1	2	328.600	16	18	0

The agglomeration schedule gives information on the cases being combined at each stage of a hierarchical clustering process. We begin with all n=20n=20 cases as individual clusters each of size 1. At the first stage, SPSS combines the two cases which are closest together based on the distance measure used (again, squared Euclidean distance was specified). We see that cases 14 and 16 were combined. Subsequently, cases 6 and 7 were combined, then 2 and 13 etc. The process continues until we have one cluster with all n=20n=20 cases within it. The right-hand columns report when a previously combined case first appeared. For example, case 14 was first combined in stage 1 and next appeared in stage 6.

Note the ‘Coefficient’ column provides a (scaled) measure of how close the objects are at each stage of the clustering. As we advance through the stages of hierarchical clustering, we have to combine clusters which are further and further apart. When there is a ‘large’ increase in the Coefficient column, at that stage we are combining quite distant clusters and so we may wish to halt the agglomeration just before this happens - otherwise we combine distant objects which we might be unwilling to consider as ‘similar’ or ‘homogeneous’.

Cluster Membership

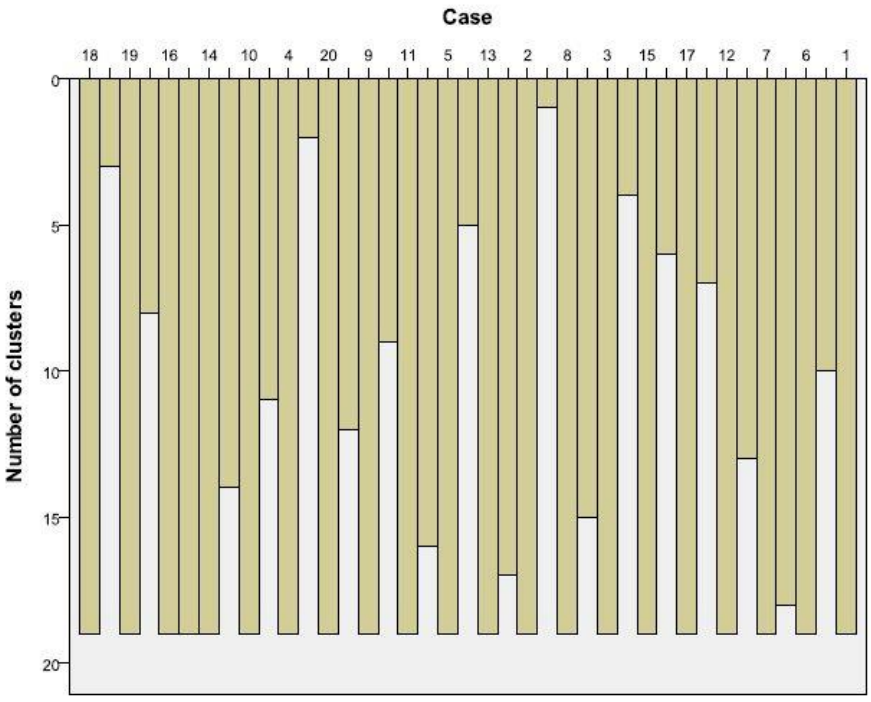
Case	4 Cluster	3 Cluster	2 Cluster
1	1	1	1
2	2	2	2
3	1	1	1
4	3	3	2
5	2	2	2
6	1	1	1
7	1	1	1
8	1	1	1
9	2	2	2
10	3	3	2
11	2	2	2
12	1	1	1
13	2	2	2
14	3	3	2
15	1	1	1
16	3	3	2
17	1	1	1
18	4	3	2
19	3	3	2
20	2	2	2

Cluster Membership

Case	4 Clusters	3 Clusters	2 Clusters
1	1	1	1
2	2	2	2
3	1	1	1
4	3	3	2
5	2	2	2
6	1	1	1
7	1	1	1
8	1	1	1
9	2	2	2
10	3	3	2
11	2	2	2
12	1	1	1
13	2	2	2
14	3	3	2
15	1	1	1
16	3	3	2
17	1	1	1
18	4	3	2
19	3	3	2
20	2	2	2

If we asked SPSS to consider a range of cluster solutions based on our desire to identify an approximate number of clusters (for example, from 2 to 4 clusters), then the cluster membership box reports to which cluster each individual case is assigned, for each type of solution. Note these results also appear as new columns in your original data matrix.

Eyeballing the four-cluster solution column, we see that if we had four clusters then only one observation (participant 18) would appear in cluster 4. We might be unwilling to entertain such a small cluster (although 1 observation in 20 might represent approximately 5% of the population, assuming the (random) sample was fairly representative of the population from which it was drawn) and we see that with three clusters this individual would be assigned to cluster 3. (Looking back at the proximity matrix, we see observation 18 is quite far from all other observations.)



The icicle plot shows how the cluster compositions for all possible cluster solutions, i.e. from 1 to 20 clusters for this example. The vertical axis details the number of clusters and the individual participants are at the top of the plot (the numbers correspond to the participant number). The icicles indicate the split between cluster solutions in terms of which observations are included in which cluster. For example, for two clusters, the shortest icicle is between participants 2 and 8. So a two-cluster solution would have the following clusters.

- Cluster 1: 1, 3, 6, 7, 8, 12, 15 and 17.
- Cluster 2: 2, 4, 5, 9, 10, 11, 13, 14, 16, 18, 19 and 20.

For three clusters, the next shortest icicle is between participants 4 and 20, hence splitting cluster 2 to give the following clusters.

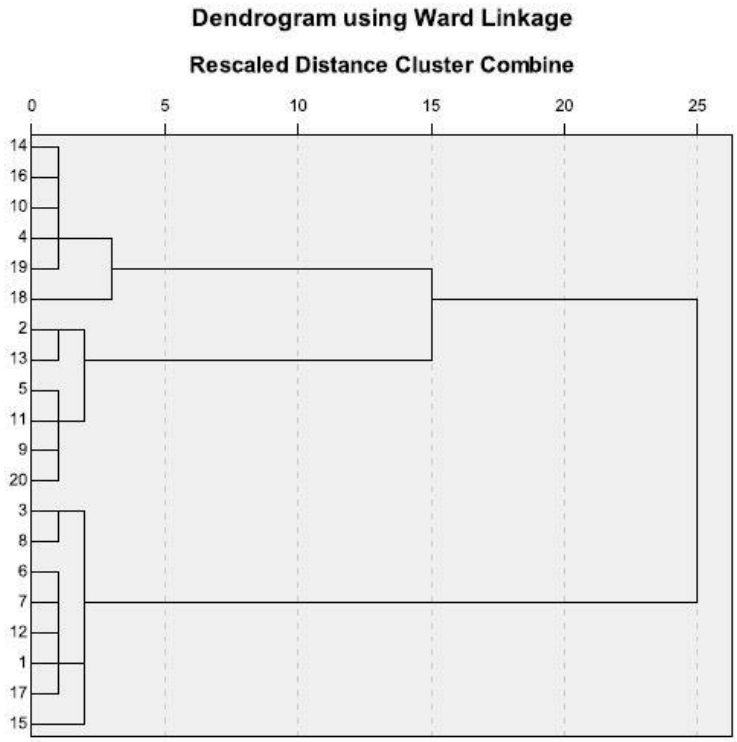
- Cluster 1: 1, 3, 6, 7, 8, 12, 15 and 17.
- Cluster 2: 2, 5, 9, 11, 13 and 20.
- Cluster 3: 4, 10, 14, 16, 18 and 19.

For four clusters, the next shortest icicle is between participants 18 and 19, hence splitting cluster 3 to give the following clusters.

- Cluster 1: 1, 3, 6, 7, 8, 12, 15 and 17.
- Cluster 2: 2, 5, 9, 11, 13 and 20.
- Cluster 3: 4, 10, 14, 16 and 19.

- Cluster 4: 18.

These results can be verified with the cluster membership output.



The dendrogram provides a convenient way to determine the appropriate number of clusters. Although the final decision is subjective, the dendrogram clearly shows the proximity when cases are combined. On the left-hand side we see the individual participants (the numbers correspond to the participant number). The horizontal axis is a rescaled distance measure showing the (rescaled) distance when combining occurs. The vertical lines in the diagram depict the combining of cases - simply trace back the cluster members to the left-hand side.

Using the above dendrogram, it seems reasonable to identify three distinct clusters.

- Cluster 1: 1, 3, 6, 7, 8, 12, 15 and 17.
- Cluster 2: 2, 5, 9, 11, 13 and 20.
- Cluster 3: 4, 10, 14, 16, 18 and 19.

Finally, we would like to profile the clusters by examining the group centroids. Given we have saved the cluster membership in the data matrix, we can obtain the means of each variable for each cluster as follows. Use **Analyze** >> **Compare Means** >> **Means**……, then move the variables used for clustering (here V_1 to V_6) into the 'Dependent List:', and move the three-cluster solution column (called 'CLU3_1') to the 'Independent List:'. Click 'OK'.

Report

Ward Method	Shopping is fun	Shopping is bad for your budget	I combine shopping with eating out	I try to get the best buys while shopping	I don't care about shopping	You can save a lot of money by comparing prices

1	**Mean**	5.75	3.63	6.00	3.13	1.88	3.88
	N	8	8	8	8	8	8
	Std. Deviation	1.035	.916	1.069	.835	.835	.641
2	**Mean**	1.67	3.00	1.83	3.50	5.50	3.33
	N	6	6	6	6	6	6
	Std. Deviation	.516	.632	.753	1.049	1.049	.816
3	**Mean**	3.50	5.83	3.33	6.00	3.50	6.00
	N	6	6	6	6	6	6
	Std. Deviation	.548	.753	.816	.632	.837	1.549
Total	**Mean**	3.85	4.10	3.95	4.10	3.45	4.35
	N	20	20	20	20	20	20
	Std. Deviation	1.899	1.410	2.012	1.518	1.761	1.496

Report

Ward Method		Shopping is fun	Shopping is bad for your budget	I combine shopping with eating out	I try to get the best buys while shopping	I don't care about shopping	You can save a lot of money by comparing prices
1	Mean	5.75	3.63	6.00	3.13	1.88	3.88
	N	8	8	8	8	8	8
	Std. Deviation	1.035	.916	1.069	.835	.835	.641
2	Mean	1.67	3.00	1.83	3.50	5.50	3.33
	N	6	6	6	6	6	6
	Std. Deviation	.516	.632	.753	1.049	1.049	.816
3	Mean	3.50	5.83	3.33	6.00	3.50	6.00
	N	6	6	6	6	6	6
	Std. Deviation	.548	.753	.816	.632	.837	1.549
Total	Mean	3.85	4.10	3.95	4.10	3.45	4.35
	N	20	20	20	20	20	20
	Std. Deviation	1.899	1.410	2.012	1.518	1.761	1.496

We see that cluster 1 members seem to be the shopaholics (high mean scores for V_1 and V_3, and a low mean score for V_5). Cluster 2 members seem to be those who hate shopping (low mean scores for V_1 and V_3, and a high mean score for V_5), while cluster 3 members seem to be price-conscious consumers (high mean scores for V_2, $V_4$and V_6).

2. Selected SPSS output follows (using the default cluster analysis options in SPSS).

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	20	0.000	0	0	12
2	18	19	0.000	0	0	3
3	17	18	0.000	0	2	10
4	3	6	0.000	0	0	15
5	14	15	0.500	0	0	9
6	8	12	1.000	0	0	11
7	5	11	1.500	0	0	13
8	1	4	2.000	0	0	14
9	10	14	2.833	0	5	14
10	16	17	4.333	0	3	13
11	8	9	5.833	6	0	18
12	2	13	7.833	1	0	16
13	5	16	11.000	7	10	16
14	1	10	14.367	8	9	17
15	3	7	18.367	4	0	17
16	2	5	31.867	12	13	19
17	1	3	46.417	14	15	18
18	1	8	65.758	17	11	19
19	1	2	138.050	18	16	0

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	20	.000	0	0	12
2	18	19	.000	0	0	3
3	17	18	.000	0	2	10
4	3	6	.000	0	0	15
5	14	15	.500	0	0	9
6	8	12	1.000	0	0	11
7	5	11	1.500	0	0	13
8	1	4	2.000	0	0	14
9	10	14	2.833	0	5	14
10	16	17	4.333	0	3	13
11	8	9	5.833	6	0	18
12	2	13	7.833	1	0	16
13	5	16	11.000	7	10	16
14	1	10	14.367	8	9	17
15	3	7	18.367	4	0	17
16	2	5	31.867	12	13	19
17	1	3	46.417	14	15	18
18	1	8	65.758	17	11	19
19	1	2	138.050	18	16	0

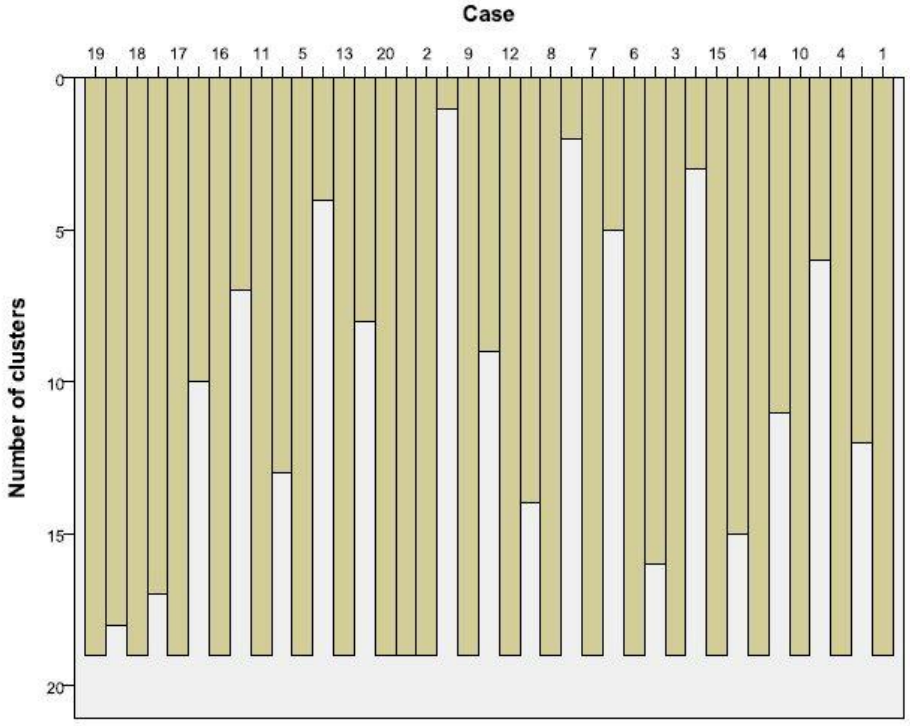
Cluster Membership

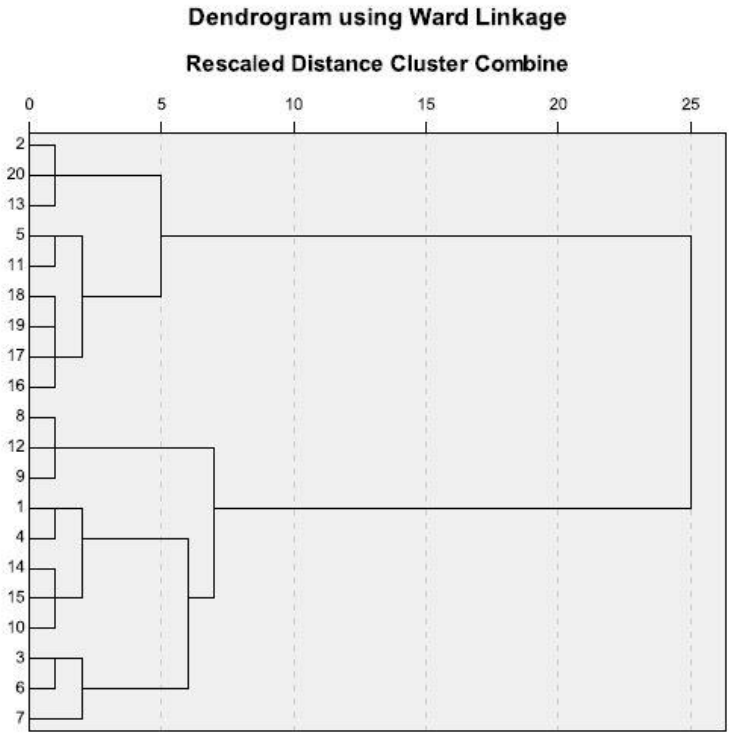
Case	4 Clusters	3 Clusters	2 Clusters
1	1	1	1
2	2	2	2
3	3	1	1
4	1	1	1
5	2	2	2
6	3	1	1
7	3	1	1
8	4	3	1
9	4	3	1
10	1	1	1
11	2	2	2
12	4	3	1
13	2	2	2
14	1	1	1
15	1	1	1
16	2	2	2
17	2	2	2

18	2	2	2
19	2	2	2
20	2	2	2

Cluster Membership

Case	4 Clusters	3 Clusters	2 Clusters
1	1	1	1
2	2	2	2
3	3	1	1
4	1	1	1
5	2	2	2
6	3	1	1
7	3	1	1
8	4	3	1
9	4	3	1
10	1	1	1
11	2	2	2
12	4	3	1
13	2	2	2
14	1	1	1
15	1	1	1
16	2	2	2
17	2	2	2
18	2	2	2
19	2	2	2
20	2	2	2





Report

Ward Method		Comfort	Style	Durability
1	Mean	4.91	4.82	5.36
	N	11	11	11
	Std. Deviation	.944	1.328	1.362
2	Mean	3.22	2.44	2.89
	N	9	9	9
	Std. Deviation	.667	.882	1.167
Total	Mean	4.15	3.75	4.25
	N	20	20	20
	Std. Deviation	1.182	1.650	1.773

Report

Ward Method		**Comfort**	**Style**	**Durability**
1	**Mean**	4.91	4.82	5.36
	N	11	11	11
	Std. Deviation	.944	1.328	1.362
2	**Mean**	3.22	2.44	2.89
	N	9	9	9
	Std. Deviation	0.667	0.882	1.167

Total	**Mean**	4.15	3.75	4.25
	N	20	20	20
	Std. Deviation	1.182	1.65	1.773

An examination of the agglomeration schedule reveals that the coefficient suddenly jumps from stage 18 to 19 (from 65.758 to 138.050). Therefore, it appears that a two-cluster solution is appropriate (the jump from stage 18 to 19 indicates we would be merging two distant clusters together). The same conclusion is reached by looking at the dendrogram (on the rescaled distance axis, the final two clusters would be merged at a distance of 25, which is far greater than the distances when smaller clusters were combined).

The cluster sizes are as follows.

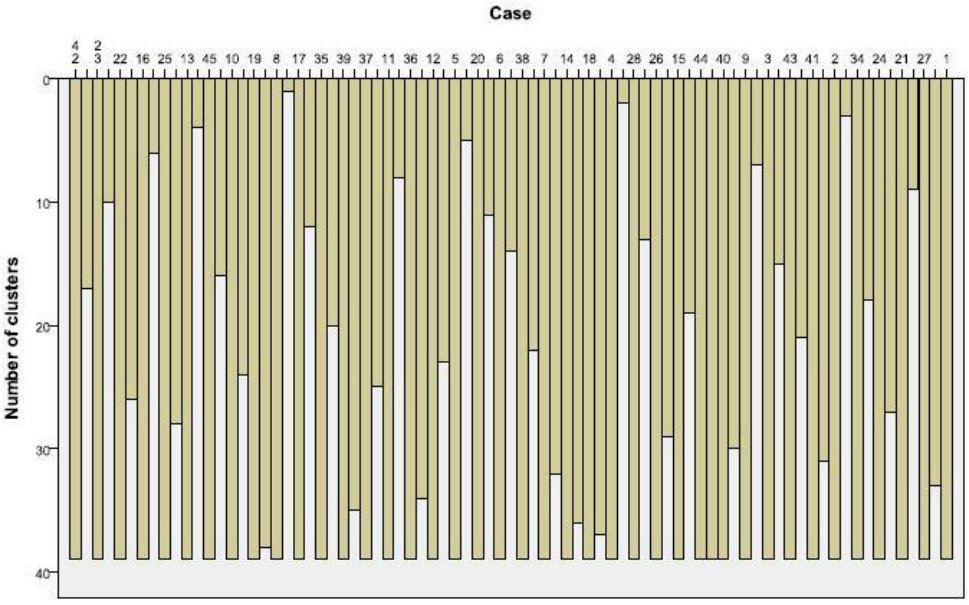
- Cluster 1: 11 (members: 1, 3, 4, 6, 7, 8, 9, 10, 12, 14 and 15).
- Cluster 2: 9 (members: 2, 5, 11, 13, 16, 17, 18, 19 and 20).

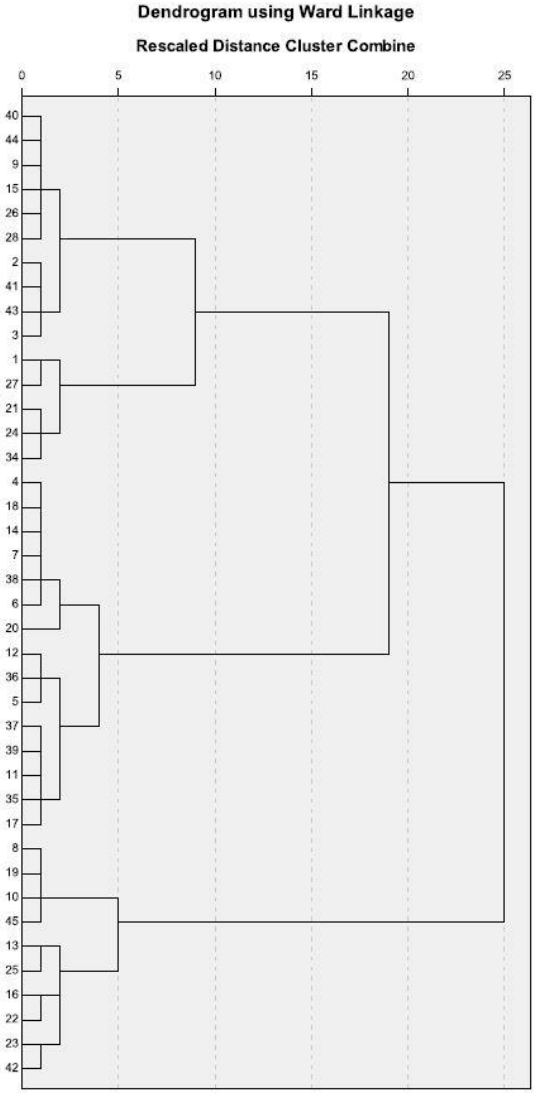
The clusters can be interpreted in terms of the values of the cluster means. Cluster 1 has high values on all three variables and may be labelled as a 'Favourable perceptions' cluster. Cluster 2 has low values on all three variables and may be labelled as an 'Unfavourable perceptions' cluster.

3. Selected SPSS output follows (using the default cluster analysis options in SPSS).

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	40	44	.500	0	0	10
2	8	19	1.000	0	0	16
3	4	18	1.500	0	0	4
4	4	14	2.333	3	0	8
5	37	39	3.333	0	0	15
6	12	36	4.333	0	0	17
7	1	27	5.333	0	0	31
8	4	7	6.750	4	0	18
9	2	41	8.250	0	0	19
10	9	40	9.750	0	1	21
11	15	26	11.250	0	0	21
12	13	25	12.750	0	0	34
13	21	24	14.250	0	0	22
14	16	22	15.750	0	0	30
15	11	37	17.417	0	5	20
16	8	10	19.583	2	0	24
17	5	12	21.917	0	6	32
18	4	38	24.367	8	0	26
19	2	43	28.200	9	0	25
20	11	35	32.283	15	0	28
21	9	15	36.383	10	11	27
22	21	34	40.883	13	0	31
23	23	42	45.883	0	0	30
24	8	45	51.717	16	0	36
25	2	3	57.883	19	0	33
26	4	6	64.683	18	0	29
27	9	28	71.917	21	0	33
28	11	17	79.967	20	0	32
29	4	20	88.538	26	0	35
30	16	23	97.288	14	23	34
31	1	21	106.688	7	22	37
32	5	11	117.930	17	28	35
33	2	9	132.096	25	27	37
34	13	16	146.346	12	30	36
35	4	5	176.133	29	32	38
36	8	13	207.633	24	34	39
37	1	2	273.533	31	33	38
38	1	4	413.067	37	35	39
39	1	8	604.400	38	36	0


Case
Number of clusters
0
10
20
30
40


Dendrogram using Ward Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25

Report

Ward Method		Awareness	Attitude	Preference	Intention	Loyalty
1	Mean	6.00	6.20	5.80	6.20	6.00
	N	5	5	5	5	5
	Std. Deviation	1.225	.447	.837	1.095	.707
2	Mean	3.50	3.20	3.80	5.60	5.60
	N	10	10	10	10	10
	Std. Deviation	1.354	.632	1.033	.843	.699
3	Mean	5.67	5.47	5.27	3.13	3.00
	N	15	15	15	15	15
	Std. Deviation	.816	1.356	1.163	.915	1.000
4	Mean	2.10	2.00	2.50	3.20	3.00
	N	10	10	10	10	10
	Std. Deviation	1.101	1.155	.850	1.398	1.633
Total	Mean	4.28	4.13	4.27	4.15	4.02
	N	40	40	40	40	40
	Std. Deviation	1.894	1.924	1.585	1.657	1.717

An examination of the agglomeration schedule reveals that the coefficient suddenly jumps from stage 36 to 37 (from 207.633 to 273.533). Therefore, it appears that a four-cluster solution is appropriate. The same conclusion is reached by looking at the dendrogram.

The cluster sizes are as follows.

- Cluster 1: 5 (members: 1, 21, 24, 27 and 34).
- Cluster 2: 10 (members: 2, 3, 9, 15, 26, 28, 40, 41, 43 and 44).
- Cluster 3: 15 (members: 4, 5, 6, 7, 11, 12, 14, 17, 18, 20, 35, 36, 37, 38 and 39).
- Cluster 4: 10 (members: 8, 10, 13, 16, 19, 22, 23, 25, 42 and 45).

There are five cases with missing values.
The clusters can be interpreted in terms of the values of the cluster means. Cluster 1 has the highest, most favourable, values on all five variables and may be labelled as ‘True loyalty’. Cluster 2 has low values on awareness, attitude and preference, but high values on intention and loyalty, and may be labelled as ‘Spurious loyalty’. Cluster 3 has high values on awareness, attitude and preference, but low values on intention and loyalty, and may be labelled as ‘Latent loyalty’. Finally, Cluster 4 has the lowest values on all five variables and may be labelled as ‘No loyalty’.

Commentary on Discussion points

1. It is paramount that the researcher seeks to validate their clustering solution to demonstrate that the clusters they are describing are ‘naturally-occurring clusters’ and not simply a consequence of how they have performed their analysis. It has been demonstrated how plausible clustering solutions can be performed on random data. If marketers build characteristics of target market segments, on clusters which do not really exist, the consequences could be disastrous.

2. The following procedures can provide adequate checks on the quality of clustering results. These are vital if managers are to appreciate what constitutes robust clustering solutions.

- Perform cluster analysis on the same data using different distance measures. Compare the results across measures to determine the stability of the solutions.
- Use different methods of clustering and compare the results.
- Split the data randomly into halves. Perform clustering separately on each half. Compare cluster centroids across the two subsamples.
- Delete variables randomly. Perform clustering based on the reduced set of variables. Compare the results with those obtained by clustering based on the entire set of variables.
- In non-hierarchical clustering, the solution may depend on the order of cases in the dataset. Make multiple runs using different orders of cases until the solution stabilises.