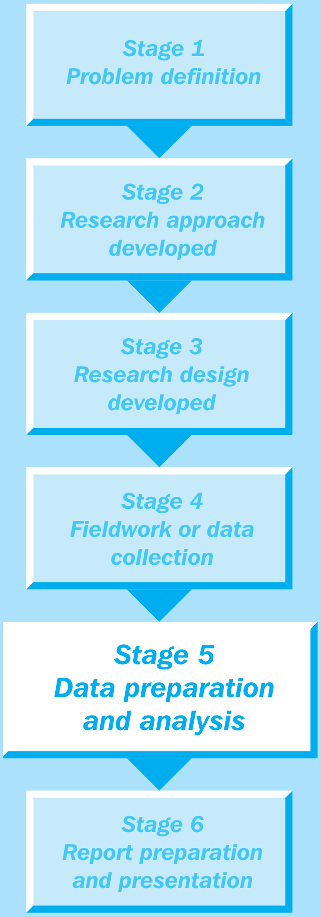
CHAPTER 23

Cluster analysis


Stage 1
Problem definition
Stage 2
Research approach developed
Stage 3
Research design developed
Stage 4
Fieldwork or data collection
Stage 5
Data preparation and analysis
Stage 6
Report preparation and presentation

Objectives

After reading this chapter, you should be able to:

1 describe the basic concept and scope of cluster analysis and its importance in marketing research;

2 discuss the statistics associated with cluster analysis;

3 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;

4 describe the purpose and methods for evaluating the quality of clustering results and assessing reliability and validity;

5 discuss the applications of non-hierarchical clustering and clustering of variables.

Cluster analysis aims to identify and classify similar entities, based upon the characteristics they possess. It helps the researcher to understand patterns of similarity and difference that reveal naturally occurring groups.

Overview

Like factor analysis (Chapter 22), cluster analysis examines an entire set of interdependent relationships. Cluster analysis makes no distinction between dependent and independent variables. Rather, interdependent relationships between the whole set of variables are examined. The primary objective of cluster analysis is to classify objects into relatively homogeneous groups based on the set of variables considered. Objects in a group are relatively similar in terms of these variables and different from objects in other groups. When used in this manner, cluster analysis is the obverse of factor analysis in that it reduces the number of objects, not the number of variables, by grouping them into a much smaller number of clusters.

This chapter describes the basic concept of cluster analysis. The steps involved in conducting cluster analysis are discussed and illustrated in the context of hierarchical clustering by using a popular computer program. Then an application of non-hierarchical clustering is presented, followed by a discussion of clustering of variables. We begin with two examples.

example
GlobalCash Project

Cluster analysis of European companies' plans over the next two years[1]

In the GlobalCash Project, respondents were clustered on the basis of the changes that respondents said their companies would be making over the next two years. The results indicated that respondents could be clustered into 20 segments. Differences among the segments were statistically tested. Thus, each segment contained respondents who were relatively homogeneous with respect to their plans. The following descriptions encapsulate four of the distinct segments.

- *Restructure through new electronic systems* represented companies whose distinctive plans involved 'making greater use of electronic banking', 'automating the treasury function', 'installing a new treasury system' and 'restructuring cash management along pan-European lines'.
- *Quality focus* represented those companies whose only significant planned change was for 'bank service quality to become a major issue'.
- *All change* represented those companies that planned changes in most areas, perhaps the most volatile of all groups. Most of this group planned to use a pan-European bank and with that change would come a restructuring of cash management on pan-European lines and installation of a new treasury system.
- *Status quo* represented the companies with the lowest amounts of planned changes. None of this group plan to have more automation in their treasury function or put their domestic banking out to tender. ■

example

Ice cream 'hot spots'[2]

In order to achieve an expanded customer base, Häagen-Dazs identified potential consumer segments that could generate additional sales. They used geodemographic techniques (as discussed in Chapter 5), which are based upon clustering consumers, using geographic, demographic and lifestyle data. Additional primary data was collected to develop an understanding of the demographic, lifestyle and behavioural characteristics of Häagen-Dazs Café users, that included frequency of purchase, time of day to visit café, day of the week and a range of other product variables. The postcodes or zip codes of respondents were also obtained. With a postcode or zip code, respondents can be assigned to one of the array of established geodemographic classifications. Häagen-Dazs compared their profile of customers to the profile of geodemographic classifications to develop a clearer picture of the types of consumer they were attracting. From this they decided which profiles of consumer or target markets they believed to hold the most potential for additional sales. ■

Both of the above examples illustrate the use of clustering to arrive at homogeneous segments for the purpose of formulating specific marketing strategies.

Basic concept

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.[3] We are concerned with clustering procedures that assign each object to one and only one cluster.[4] Figure 23.1 shows an ideal clustering situation in which the clusters are distinctly separated on two variables: quality consciousness (variable 1) and price sensitivity (variable 2). Note that each consumer falls into one cluster and there are no overlapping areas. Figure 23.2, on the other hand, presents a clustering situation more likely to be encountered in practice. In Figure 23.2, the boundaries for some of the clusters are not clear cut, and the classification of some consumers is not obvious, because many of them could be grouped into one cluster or another.

Both cluster analysis and discriminant analysis are concerned with classification. Discriminant analysis, however, 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*.[5] Cluster analysis has been used in marketing for a variety of purposes, including the following:[6]

- *Segmenting the market.* For example, consumers may be clustered on the basis of benefits sought from the purchase of a product. Each cluster would consist of consumers who are relatively homogeneous in terms of the benefits they seek.[7] This approach is called benefit segmentation.

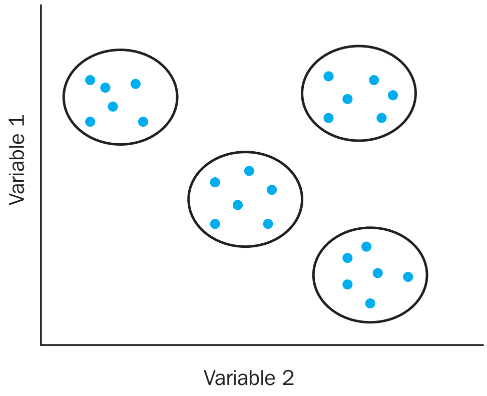


Figure 23.1
An ideal clustering solution

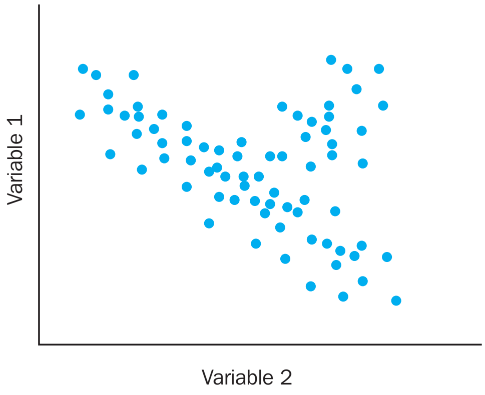


Figure 23.2
A practical clustering solution

- *Understanding buyer behaviours.* Cluster analysis can be used to identify homogeneous groups of buyers. Then the buying behaviour of each group may be examined separately, as happened in another area of the GlobalCash Project. Respondents were clustered on the basis of choice criteria used in selecting a bank.
- *Identifying new product opportunities.* By clustering brands and products, competitive sets within the market can be determined. Brands in the same cluster compete more fiercely with each other than with brands in other clusters. A firm can examine its current offerings compared with those of its competitors to identify potential new product opportunities.
- *Selecting test markets.* By grouping cities into homogeneous clusters, it is possible to select comparable cities to test various marketing strategies.
- *Reducing data.* Cluster analysis can be used as a general data reduction tool to develop clusters or subgroups of data that are more manageable than individual observations. Subsequent multivariate analysis is conducted on the clusters rather than on the individual observations. For example, to describe differences in consumers' product usage behaviour, the consumers may first be clustered into groups. The differences among the groups may then be examined using multiple discriminant analysis.

Before discussing the statistics associated with cluster analysis, it should be mentioned that most clustering methods are relatively simple procedures that are not supported by an extensive body of statistical reasoning. Rather, most clustering methods are heuristics, which are based on algorithms. Thus, cluster analysis contrasts sharply with analysis of variance, regression, discriminant analysis and factor analysis, which are based upon an extensive body of statistical reasoning. Although many clustering methods have important statistical properties, the fundamental simplicity of these methods needs to be recognised.[8] The following statistics and concepts are 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 that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. The dendrogram is read from left to right. Figure 23.8 later in this chapter is a dendrogram.

Distances between cluster centres. These distances indicate how separated the individual pairs of clusters are. Clusters that 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 bottom to top. Figure 23.7 later in this chapter is an icicle diagram.

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

Conducting cluster analysis

The steps involved in conducting cluster analysis are listed in Figure 23.3. The first step is to formulate the clustering problem by defining the variables on which the clustering will be based. Then, an appropriate distance measure must be selected. The distance measure determines how similar or dissimilar the objects being clustered are. Several clustering procedures have been developed, and the researcher should select one that is appropriate for the problem at hand. Deciding on the number of clusters requires judgement on the part of the researcher. The derived clusters should be interpreted in terms of the variables used to cluster them and profiled in terms of additional salient variables. Finally, the researcher must assess the validity of the clustering process.

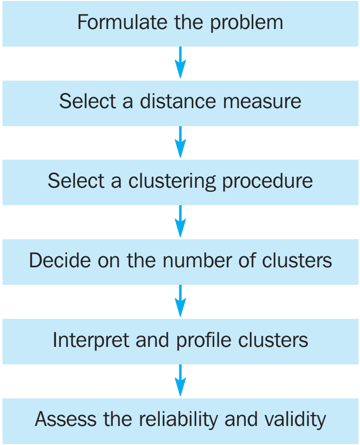

Formulate the problem
Select a distance measure
Select a clustering procedure
Decide on the number of clusters
Interpret and profile clusters
Assess the reliability and validity

Figure 23.3
Conducting cluster analysis

Formulate the problem

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 that are relevant to the marketing research problem. The variables should be selected based on past research, theory or a consideration of the hypotheses being developed or tested. If cluster analysis is used as an exploratory approach, the researcher naturally exercises their judgement and intuition.

To illustrate, we consider a clustering of consumers based on attitudes towards shopping. Based on past research, six attitudinal variables were identified as being the most relevant to the marketing research problem. Consumers were asked to express their degree of agreement with the following statements on a seven-point scale (1 = disagree, 7 = 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.

Data obtained from a pre-test sample of 20 respondents are shown in Table 23.1. Note that, in practice, clustering is done on much larger samples of 100 or more. A small sample size has been used to illustrate the clustering process.

Table 23.1 Attitudinal data for clustering

Case number	V_1	V_2	V_3	V_4	V_5	V_6
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

Select a distance measure

Because the objective of clustering is to group similar objects together, some measure is needed to assess how similar or different the objects are. The most common approach is to measure similarity in terms of distance between pairs of objects. Objects with smaller distances between them are more similar to each other than are those at larger distances. There are several ways to compute the distance between two objects.[9]

Euclidean distance
The square root of the sum of the squared differences in values for each variable.

The most commonly used measure of similarity is the **euclidean distance** or its square.[10] 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. For our example, we use the squared euclidean distance.

If the variables are measured in vastly different units, the clustering solution will be influenced by the units of measurement. In a supermarket shopping study, attitudinal variables may be measured on a nine-point Likert-type scale; patronage, in terms of frequency of visits per month and the amount spent; and brand loyalty, in terms of percentage of grocery shopping expenditure allocated to the favourite supermarket. In these cases, before clustering respondents, we must standardise the data by rescaling each variable to have a mean of zero and a standard deviation of unity. Although standardisation can remove the influence of the unit of measurement, it can also reduce the differences between groups on variables that may best discriminate groups or clusters. It is also desirable to eliminate outliers (cases with atypical values).[11]

Use of different distance measures may lead to different clustering results. Hence, it is advisable to use different measures and to compare the results. Having selected a distance or similarity measure, we can next select a clustering procedure.

Hierarchical clustering
A clustering procedure characterised by the development of a hierarchy or treelike structure.

Agglomerative clustering
A hierarchical clustering procedure where each object starts out in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters.

Divisive clustering
A hierarchical clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

Linkage methods
Agglomerative methods of hierarchical clustering that cluster objects based on a computation of the distance between them.

Single linkage
A linkage method based on minimum distance or the nearest neighbour rule.

Complete linkage
A linkage method that is based on maximum distance or the farthest neighbour approach.

Select a clustering procedure

Figure 23.4 is a classification of clustering procedures.

Clustering procedures can be hierarchical or non-hierarchical. **Hierarchical clustering** is characterised by the development of a hierarchy or treelike 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 bigger and bigger 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 marketing research. They consist of linkage methods, error sums of squares or variance methods, and centroid methods. **Linkage methods** include single linkage, complete linkage and average linkage. The **single linkage** method is based on minimum distance or the nearest neighbour rule. The first two objects clustered are those that have the smallest distance between them. The next shortest distance is identified, and either the third object is clustered with the first two or a new two-object cluster is formed. At every stage, the distance between two clusters is the distance between their two closest points (see Figure 23.5). Two clusters are merged at any stage by the single shortest link between them. This process is continued until all objects are in one cluster. The single linkage method does not work well when the clusters are poorly defined. 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 (see Figure 23.5). The average linkage

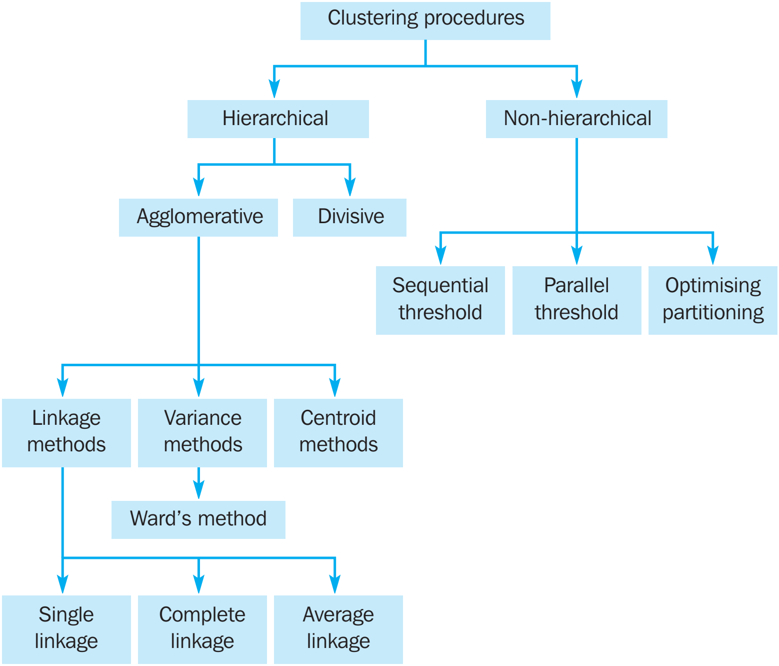


Figure 23.4
A classification of clustering procedures

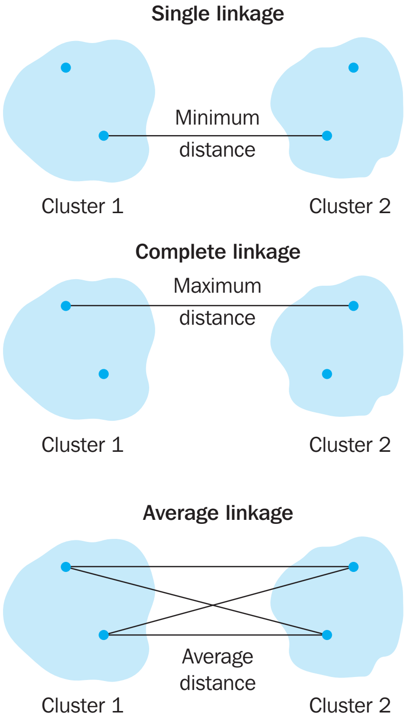


Figure 23.5
Linkage methods of clustering

Variance method
An agglomerative method of hierarchical clustering in which clusters are generated to minimise the within-cluster variance.

Ward's procedure
A variance method in which the squared euclidean distance to the cluster means is minimised.

Centroid method
A variance method of hierarchical clustering in which the distance between two clusters is the distance between their centroids (means for all the variables).

method works similarly. In this method, however, 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 23.5). As can be seen, the average linkage method uses information on all pairs of distances, not merely the minimum or maximum distances. For this reason, it is usually preferred to the single and complete linkage methods.

The **variance methods** attempt 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. Then, for each object, the squared euclidean distance to the cluster means is calculated (Figure 23.6), and 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), as shown in Figure 23.6. Every time objects are

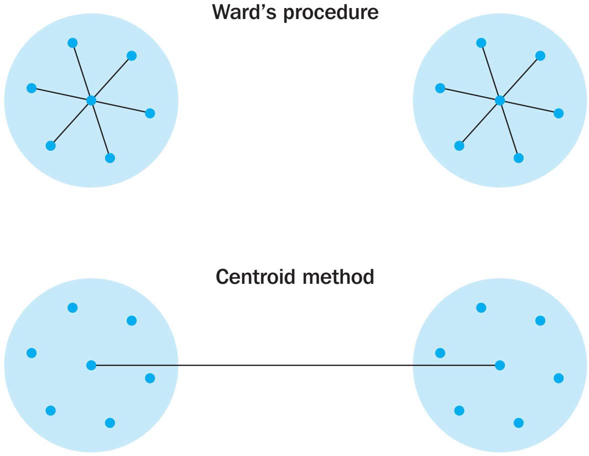


Figure 23.6
Other agglomerative clustering methods

grouped, a new centroid is computed. Of the hierarchical methods, the average linkage method and Ward's procedure have been shown to perform better than the other procedures.[12]

Non-hierarchical clustering
A procedure that first assigns or determines a cluster centre and then groups all objects within a pre-specified threshold value from the centre.

Parallel threshold method
A non-hierarchical clustering method that specifies several cluster centres at once. All objects that are within a pre-specified threshold value from the centre are grouped together.

Optimising partitioning method
A non-hierarchical clustering method that allows for later reassignment of objects to clusters to optimise an overall criterion.

The second type of clustering procedures, the **non-hierarchical clustering** methods, are frequently referred to as *k*-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 prespecified threshold value from the centre are grouped together. A new cluster centre or seed is then 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.

Two major disadvantages of the non-hierarchical procedures are that the number of clusters must be prespecified and that the selection of cluster centres is arbitrary. Furthermore, the clustering results may depend on how the centres are selected. Many non-hierarchical programs select the first k cases (k = number of clusters) without missing values as initial cluster centres. Thus, the clustering results may depend on the order of observations in the data. Yet non-hierarchical clustering is faster than hierarchical methods and has merit when the number of objects or observations is large. 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.[13]

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

We will use Ward's procedure to illustrate hierarchical clustering. The output obtained by clustering the data of Table 23.1 is given in Table 23.2. Useful information is contained in the agglomeration schedule, which shows the number of cases or clusters being combined at each stage. The first line represents stage 1, with 19 clusters. Respondents 14 and 16 are combined at this stage, as shown in the columns labelled 'Clusters combined'. The squared euclidean distance between these two respondents is given under the column labelled 'Coefficient'. The column entitled 'Stage cluster first appears' indicates the stage at which a cluster is first formed. To illustrate, an entry of 1 at stage 7 indicates that respondent 14 was first grouped at stage 1. The last column, 'Next stage', indicates the stage at which another case (respondent) or cluster is combined with this one. Because the number in the first line of the last column is 7, we see that, at stage 7, respondent 10 is combined with 14 and 16 to form a single cluster. Similarly, the second line represents stage 2 with 18 clusters. In stage 2, respondents 2 and 13 are grouped together.

Another important part of the output is contained in the icicle plot given in Figure 23.7. The columns correspond to the objects being clustered; in this case, they are the respondents labelled 1 to 20. The rows correspond to the number of clusters. This figure is read from bottom to top. At first, all cases are considered as individual clusters. Since there are 20 respondents, there are 20 initial clusters. At the first step, the two closest objects are combined, resulting in 19 clusters. The last line of Figure 23.7 shows these 19 clusters. The two cases, respondents 14 and 16, that have been combined at this stage have no blank space separating them. Row number 18 corresponds

Table 23.2 **Results of hierarchical clustering**

Agglomeration schedule using Ward's procedure

	Clusters combined			*Stage cluster first appears*		
Stage	*Cluster 1*	*Cluster 2*	*Coefficient*	*Cluster 1*	*Cluster 2*	*Next stage*
1	14	16	1.000000	0	0	7
2	2	13	2.500000	0	0	15
3	7	12	4.000000	0	0	10
4	5	11	5.500000	0	0	11
5	3	8	7.000000	0	0	16
6	1	6	8.500000	0	0	10
7	10	14	10.166667	0	1	9
8	9	20	12.666667	0	0	11
9	4	10	15.250000	0	7	12
10	1	7	18.250000	6	3	13
11	5	9	22.750000	4	8	15
12	4	19	27.500000	9	0	17
13	1	17	32.700001	10	0	14
14	1	15	40.500000	13	0	16
15	2	5	51.000000	2	11	18
16	1	3	63.125000	14	5	19
17	4	18	78.291664	12	0	18
18	2	4	171.291656	15	17	19
19	1	2	330.450012	16	18	0

Cluster membership of cases using Ward's procedure

	Number of clusters		
Label case	*4*	*3*	*2*
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

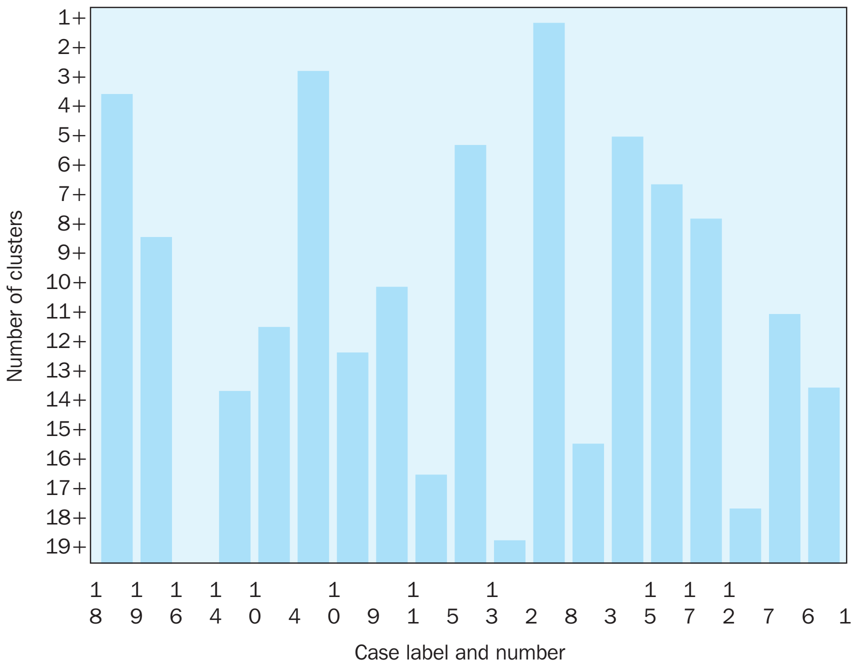


Figure 23.7
Vertical icicle plot using Ward's procedure

to the next stage, with 18 clusters. At this stage, respondents 2 and 13 are grouped together. Thus, at this stage there are 18 clusters; 16 of them consist of individual respondents, and two contain two respondents each. Each subsequent step leads to the formation of a new cluster in one of three ways: (1) two individual cases are grouped together, (2) a case is joined to an already existing cluster, or (3) two clusters are grouped together.

Another graphic device that is useful in displaying clustering results is the dendrogram (see Figure 23.8). The dendrogram is read from left to right. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. Because many distances in the early stages are of similar magnitude, it is difficult to tell the sequence in which some of the early clusters are formed. It is clear, however, that in the last two stages, the distances

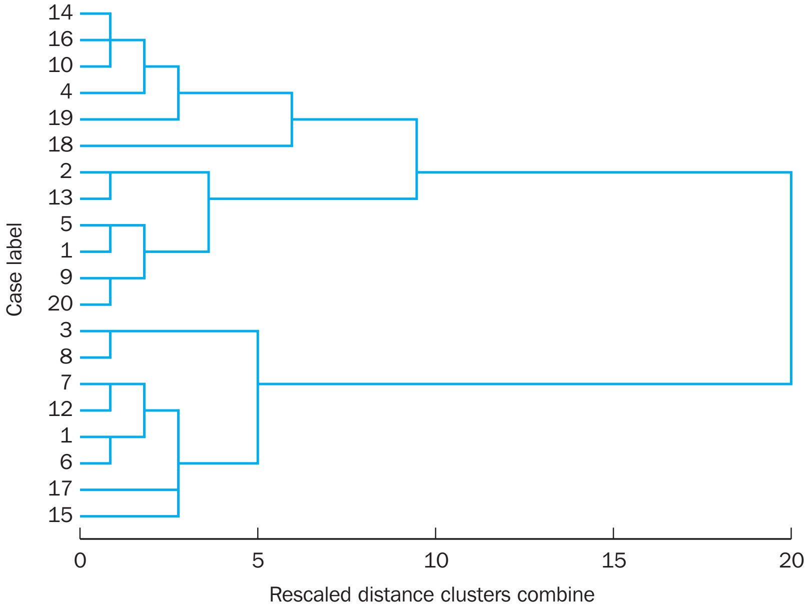


Figure 23.8
Dendrogram using Ward's procedure

at which the clusters are being combined are large. This information is useful in deciding on the number of clusters.

It is also possible to obtain information on cluster membership of cases if the number of clusters is specified. Although this information can be discerned from the icicle plot, a tabular display is helpful. Table 23.2 contains the cluster membership for the cases, depending on whether the final solution contains two, three or four clusters. Information of this type can be obtained for any number of clusters and is useful for deciding on the number of clusters.

Decide on the number of clusters

A major issue in cluster analysis is deciding on the number of clusters. Although there are no hard and fast rules, some guidelines are available.

1 Theoretical, conceptual or practical considerations may suggest a certain number of clusters. For example, if the purpose of clustering is to identify market segments, management may want a particular number of clusters.

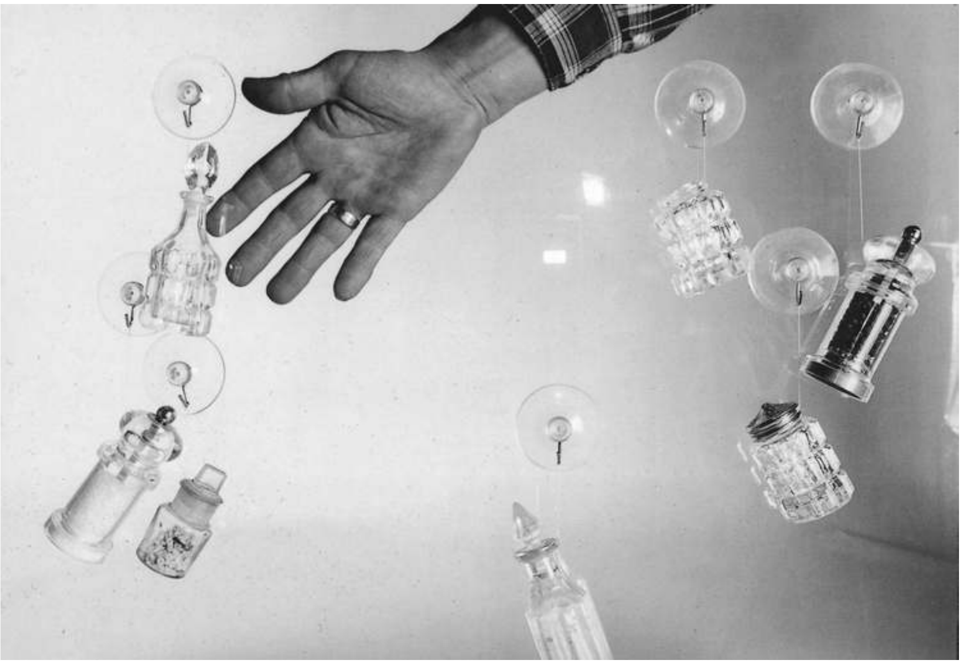
2 In hierarchical clustering, the distances at which clusters are combined can be used as criteria. This information can be obtained from the agglomeration schedule or from the dendrogram. In our case, we see from the agglomeration schedule in Table 23.2 that the value in the 'Coefficient' column suddenly more than doubles between stages 17 and 18. Likewise, at the last two stages of the dendrogram in Figure 23.8, the clusters are being combined at large distances. Therefore, it appears that a three-cluster solution is appropriate.

3 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. Increasing the number of clusters beyond this point is usually not worthwhile.

4 The relative sizes of the clusters should be meaningful. In Table 23.2, by making a simple frequency count of cluster membership, we see that a three-cluster solution results in clusters with eight, six and six elements. If we go to a four-cluster solution, however, the sizes of the clusters are eight, six, five and one. It is not meaningful to have a cluster with only one case, so a three-cluster solution is preferable in this situation.

Interpret and profile clusters

Interpreting and profiling clusters involves examining the cluster centroids. The centroids represent the mean values of the objects contained in the cluster on each of the variables. The centroids enable us to describe each cluster by assigning it a name or label. If the clustering program does not print this information, it may be obtained through discriminant analysis. Table 23.3 gives the centroids or mean values for each cluster in our example. Cluster 1 has relatively high values on variables V_1 (Shopping is fun) and V_3 (I combine shopping with eating out). It also has a low value on V_5 (I don't care about shopping). Hence cluster 1 could be labelled 'fun-loving and concerned shoppers'. This cluster consists of cases 1, 3, 6, 7, 8, 12, 15 and 17. Cluster 2 is just the opposite, with low values on V_1 and V_3 and a high value on V_5 and this cluster could be labelled 'apathetic shoppers'. Members of cluster 2 are cases 2, 5, 9, 11, 13 and 20. Cluster 3 has high values on V_2 (Shopping is bad for your budget), V_4 (I try to get the best buys while shopping) and V_6 (You can save a lot of money by comparing prices). Thus, this cluster could be labelled 'economical shoppers'. Cluster 3 is composed of cases 4, 10, 14, 16, 18 and 19. It is often helpful to profile the clusters in terms of variables that were not used for clustering, such as demographic, psychographic, product usage, media usage or other variables. For example, the clusters may have been derived based on benefits sought. Further profiling may be done in terms



Interpreting and profiling clusters involves examining the cluster centroids.

of demographic and psychographic variables to target marketing efforts for each cluster. The variables that significantly differentiate between clusters can be identified via discriminant analysis and one-way analysis of variance.

Table 23.3 Cluster centroids

Cluster number	*Means of variables*					
	V_1	V_2	V_3	V_4	V_5	V_6
1	5.750	3.625	6.000	3.125	1.750	3.875
2	1.667	3.000	1.833	3.500	5.500	3.333
3	3.500	5.833	3.333	6.000	3.500	6.000

Assess the reliability and validity

Given the several judgements entailed in cluster analysis, no clustering solution should be accepted without some assessment of its reliability and validity. Formal procedures for assessing the reliability and validity of clustering solutions are complex and not fully defensible.[14] Hence, we omit them here. The following procedures, however, provide adequate checks on the quality of clustering results.

1. Perform cluster analysis on the same data using different distance measures. Compare the results across measures to determine the stability of the solutions.
2. Use different methods of clustering and compare the results.
3. Split the data randomly into halves. Perform clustering separately on each half. Compare cluster centroids across the two sub-samples.
4. 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.
5. In non-hierarchical clustering, the solution may depend on the order of cases in the data set. Make multiple runs using different order of cases until the solution stabilises.

We further illustrate hierarchical clustering with a study of differences in marketing strategy among British, Japanese and US firms.

example

It is a small world[15]

Data for a study of British, Japanese and US competitors were obtained from detailed personal interviews with chief executives and top marketing decision makers for defined product groups in 90 companies. To control for market differences, the methodology was based upon matching 30 British companies with their major Japanese and American competitors in the British market. The study involved 30 triads of companies, each composed of a British, US and Japanese business that competed directly with one another.

Most of the data on the characteristics of the companies' performance, strategy and organisation were collected on five-point semantic differential scales. The first stage of the analysis involved factor analysis of variables describing the firms' strategies and marketing activities. The factor scores were used to identify groups of similar companies using Ward's hierarchical clustering routine. A six-cluster solution was developed.

Strategic clusters

Cluster	*I*	*II*	*III*	*IV*	*V*	*VI*
Name	*Innovators*	*Quality marketeers*	*Price promoters*	*Product marketeers*	*Mature marketeers*	*Aggressive pushers*
Size	22	11	14	13	13	17
Successful (%)	55	100	36	38	77	41
Nationality (%)						
British	23	18	64	38	31	29
Japanese	59	46	22	31	15	18
American	18	36	14	31	54	53

Membership in the six clusters was then interpreted against the original performance, strategy and organisational variables. All the clusters contained some successful companies, although some contained significantly more than others. The clusters lent support to the hypothesis that successful companies were similar irrespective of nationality, since British, Japanese and US companies were found in all the clusters. There was, however, a preponderance of Japanese companies in the more successful clusters and a predominance of British companies in the two least successful clusters. Apparently, Japanese companies do not deploy strategies that are unique to them; rather, more of them pursue strategies that work effectively in the British market.

The findings indicate that there are generic strategies that describe successful companies irrespective of their industry. Three successful strategies can be identified. The first is the Quality Marketing strategy. These companies have strengths in marketing and research and development. They concentrate their technical developments on achieving high quality rather than pure innovation. These companies are characterised by entrepreneurial organisations, long-range planning and a well-communicated sense of mission. The second generic strategy is that of the Innovators who are weaker on advanced research and development but are entrepreneurial and driven by a quest for innovation. The last successful group are the Mature Marketers, who are highly profit oriented and have in-depth marketing skills. All three appear to consist of highly marketing-oriented businesses. ■

Applications of non-hierarchical clustering

We illustrate the non-hierarchical procedure using the data in Table 23.1 and an optimising partitioning method. Based on the results of hierarchical clustering, a three-cluster solution was prespecified. The results are presented in Table 23.4.

The initial cluster centres are the values of the first three cases. The classification cluster centres are interim centres used for the assignment of cases. Each case is assigned to the nearest classification cluster centre. The classification centres are

Table 23.4 Results of non-hierarchical clustering

Initial cluster centres

Cluster	V_1	V_2	V_3	V_4	V_5	V_6
1	4.0000	6.0000	3.0000	7.0000	2.0000	7.0000
2	2.0000	3.0000	2.0000	4.0000	7.0000	2.0000
3	7.0000	2.0000	6.0000	4.0000	1.0000	3.0000

Classification cluster centres

Cluster	V_1	V_2	V_3	V_4	V_5	V_6
1	3.8135	5.8992	3.2522	6.4891	2.5149	6.6957
2	1.8507	3.0234	1.8327	3.7864	6.4436	2.5056
3	6.3558	2.8356	6.1576	3.6736	1.3047	3.2010

Case listing of cluster membership

Case ID	*Cluster*	*Distance*
1	3	1.780
2	2	2.254
3	3	1.174
4	1	1.882
5	2	2.525
6	3	2.340
7	3	1.862
8	3	1.410
9	2	1.843
10	1	2.112
11	2	1.923
12	3	2.400
13	2	3.382
14	1	1.772
15	3	3.605
16	1	2.137
17	3	3.760
18	1	4.421
19	1	0.853
20	2	0.813

Final cluster centres

Cluster	V_1	V_2	V_3	V_4	V_5	V_6
1	3.5000	5.8333	3.3333	6.0000	3.5000	6.0000
2	1.6667	3.0000	1.8333	3.5000	5.5000	3.3333
3	5.7500	3.6250	6.0000	3.1250	1.7500	3.8750

Table 23.4 Continued

Distances between final cluster centres

Cluster	*1*	*2*	*3*
1	0.0000		
2	5.5678	0.0000	
3	5.7353	6.9944	0.0000

Analysis of variance

Variable	*Cluster MS*	*df*	*Error MS*	*df*	*F*	*p*
V_1	29.1083	2	0.6078	17.0	47.8879	0.000
V_2	13.5458	2	0.6299	17.0	21.5047	0.000
V_3	31.3917	2	0.8333	17.0	37.6700	0.000
V_4	15.7125	2	0.7279	17.0	21.5848	0.000
V_5	24.1500	2	0.7353	17.0	32.8440	0.000
V_6	12.1708	2	1.0711	17.0	11.3632	0.001

Number of cases in each cluster

Cluster	*Unweighted cases*	*Weighted cases*
1	6.0	6.0
2	6.0	6.0
3	8.0	8.0
Missing	0.0	
Total	20.0	20.0

updated until the stopping criteria are reached. The final cluster centres represent the variable means for the cases in the final clusters.

Table 23.4 also displays cluster membership and the distance between each case and its classification centre. Note that the cluster memberships given in Table 23.2 (hierarchical clustering) and Table 23.4 (non-hierarchical clustering) are identical. (Cluster 1 of Table 23.2 is labelled cluster 3 in Table 23.4, and cluster 3 of Table 23.2 is labelled cluster 1 in Table 23.4.) The distances between the final cluster centres indicate that the pairs of clusters are well separated. The univariate *F* test for each clustering variable is presented. These *F* tests are only descriptive. Because the cases or objects are systematically assigned to clusters to maximise differences on the clustering variables, the resulting probabilities should not be interpreted as testing the null hypothesis of no differences among clusters.

The following example of hospital choice further illustrates non-hierarchical clustering.

example

Segmentation with surgical precision[16]

Cluster analysis was used to classify and segment respondents, based upon their preferences for hospitals that provide in-patient care. The clustering was based on the reasons respondents gave for preferring a particular hospital. The demographic profiles of the grouped respondents were compared to learn whether the segments could be identified more efficiently.

Because different individuals perceive scales of importance differently, each individual's ratings were normalised before clustering. The results indicated that the respondents could be best classified into four clusters. The cross-validation procedure for cluster analysis was run twice, on halves of the total sample.

As expected, the four groups differed substantially by their distributions and average responses to the reasons for their hospital preferences. The names assigned to the four groups reflected the demographic characteristics and reasons for hospital preferences: '*old-fashioned*', '*affluent*', '*value conscious*', and '*professional want-it-alls*'. ■

Clustering variables

Sometimes cluster analysis is also used for clustering variables to identify homogeneous groups. In this instance, the units used for analysis are the variables, and the distance measures are computed for all pairs of variables. For example, the correlation coefficient, either the absolute value or with the sign, can be used as a measure of similarity (the opposite of distance) between variables.

Hierarchical clustering of variables can aid in the identification of unique variables, or variables that 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. A given number of cluster components does not generally explain as much variance as the same number of principal components, however. Why, then, should the clustering of variables be used? Cluster components are usually easier to interpret than the principal components, even if the latter are rotated.[17] We illustrate the clustering of variables with an example from advertising research.

example

Feelings – nothing more than feelings[18]

A study was conducted to identify feelings that are precipitated by advertising. A total of 655 feelings were reduced to a set of 180 that were judged by respondents to be most likely to be stimulated by advertising. This group was clustered on the basis of judgements of similarity between feelings resulting in 31 feelings clusters. These were divided into 16 positive and 15 negative clusters, as shown in the table.

Positive feelings		Negative feelings	
1	Playful/childish	1	Affraid
2	Friendly	2	Bad/sick
3	Humorous	3	Confused
4	Delighted	4	Indifferent
5	Interested	5	Bored
6	Strong/confident	6	Sad
7	Warm/tender	7	Anxious
8	Relaxed	8	Helpless/timid
9	Energetic/impulsive	9	Ugly/stupid
10	Eager/excited	10	Pity/deceived
11	Contemplative	11	Mad
12	Proud	12	Disagreeable
13	Persuaded/expectant	13	Disgusted
14	Vigorous/challenged	14	Irritated
15	Amazed	15	Moody/frustrated
16	Set/informed		

Thus, 655 feelings responses to advertising were reduced to a core set of 31 feelings. In this way, advertisers now have a manageable set of feelings for understanding and measuring responses to advertising. When measured, these feelings can provide information on a commercial's ability to persuade target consumers ■

Internet and computer applications

SPSS[19]

The main program for hierarchical clustering of objects or cases is CLUSTER. Different distance measures can be computed, and all the hierarchical clustering procedures discussed here are available. For non-hierarchical clustering, the QUICK CLUSTER program can be used. This program is particularly helpful for clustering a large number of cases. All the default options will result in a *k*-means clustering. To cluster variables, the distance measures should be computed across variables using the PROXIMITIES program. This proximity matrix can be read into CLUSTER to obtain a grouping of the variables.

SAS

The CLUSTER program can be used for the hierarchical clustering of cases or objects. All the clustering procedures discussed here are available, as well as some additional ones. Non-hierarchical clustering of cases or objects can be accomplished using FASTCLUS. For clustering of variables, the VARCLUS program can be used. Dendrograms are not automatically computed but can be obtained using the TREE program.

Minitab

Cluster analysis can be accessed in the Multivariate>Cluster observation function. Also available are Clustering of Variables and Cluster *K*-Means.

Excel

At the time of writing, cluster analysis was not available.

Summary

Cluster analysis is used for classifying objects or cases, and sometimes variables, into relatively homogeneous groups. The groups or clusters are suggested by the data and are not defined *a priori*.

The variables on which the clustering is based should be selected based on past research, theory, the hypotheses being tested, or the judgement of the researcher. An appropriate measure of distance or similarity should be selected. The most commonly used measure is the euclidean distance or its square.

Clustering procedures may be hierarchical or non-hierarchical. Hierarchical clustering is characterised by the development of a hierarchy or treelike structure. Hierarchical methods can be agglomerative or divisive. Agglomerative methods consist of linkage methods, variance methods and centroid methods. Linkage methods are composed of single linkage, complete linkage and average linkage. A commonly used variance method is the Ward's procedure. The non-hierarchical methods are frequently referred to as *k*-means clustering. These methods can be classified as sequential threshold, parallel threshold and optimising partitioning. Hierarchical and non-hierarchical methods can be used in tandem. The choice of a clustering procedure and the choice of a distance measure are interrelated.

The number of clusters may be based on theoretical, conceptual or practical considerations. In hierarchical clustering, the distances at which the clusters are being combined is an important criterion. The relative sizes of the clusters should be meaningful. The clusters should be interpreted in terms of cluster centroids. It is often helpful to profile the clusters in terms of variables that were not used for clustering. The reliability and validity of the clustering solutions may be assessed in different ways.

Questions

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

2 What is a 'cluster'?

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

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

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

6 Present a classification of clustering procedures.

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

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

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

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

11 What is involved in the interpretation of clusters?

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

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

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

15 How is cluster analysis used to group variables?

Notes

[1] Birks, D.F. and Birts, A.N., 'Cash management market segmentation', in Birks, D.F. (ed.), *Global Cash Management in Europe* (Basingstoke, Macmillan, 1998), 83–109.

[2] Stuart, L., 'Häagen-Dazs aims to scoop a larger share', *Marketing Week* 19(46/2) (21 February 1997), 26.

[3] For applications of cluster analysis, see Kale, S.H., 'Grouping euroconsumers: a culture-based clustering approach', *Journal of International Marketing* 3(3) (1995), 35–48; Day, G.S. and Nedungali, P., 'Managerial representation of competitive advantage', *Journal of Marketing* 58 (April 1994), 31–44.

[4] Overlapping clustering methods that permit an object to be grouped into more than one cluster are also available. See Chaturvedi, A., Carroll, J.D., Green, P.E. and Rotondo, J.A., 'A feature based approach to market segmentation via overlapping k-centroids clustering', *Journal of Marketing Research* 34 (August 1997), 370–7.

[5] Excellent discussions on the various aspects of cluster analysis may be found in Aldenderfer, M.S. and Blashfield, R.K., *Cluster Analysis* (Beverly Hills. CA: Sage, 1984); Everitt, B., *Cluster Analysis*, 3rd edn (New York: Halsted Press, 1993); and Romsburg, H.C., *Cluster Analysis for Researchers* (Melbourne: Krieger Publishing, 1990).

[6] Douglas, V., 'Questionnaires too long? Try variable clustering', *Marketing News* 29(5) (27 February 1995), 38; Punj, G. and Stewart, D., 'Cluster analysis in marketing research: review and suggestions for application', *Journal of Marketing Research* 20 (May 1983), 134–48.

7 For use of cluster analysis for segmentation, see Peterson, M. and Malhotra, N.K., 'Comparative marketing measures of societal quality of life: substantive dimensions in 186 countries', *Journal of Macromarketing* 17(1) (Spring 1997), 25–38; Chang, T.-Z. and Chen, S.-J., 'Benefit segmentation: a useful tool for financial investment services', *Journal of Professional Services Marketing* 12(2) (1995), 69–80; 'Using cluster analysis for segmentation', *Sawtooth News* 10 (Winter 1994–95), 6–7.

8 Everitt, B., *Cluster Analysis*, 3rd edn (New York: Halstead Press, 1993).

9 For a detailed discussion on the different measures of similarity, and formulas for computing them, see Chepoi, V. and Dragan, F., 'Computing a median point of a simple rectilinear polygon', *Information Processing Letters* 49(6) (22 March 1994), 281–5; Romsburg, H.C., *Cluster Analysis for Researchers* (Belmont, CA: Lifetime Learning Publications, 1984).

10 Hirata, T., 'A unified linear-time algorithm for computing distance maps', *Information Processing Letters* 58(3) (13 May 1996) 129–33; Hair Jr, J.E., Anderson, R.E., Tatham, R.L. and Black, W.C., *Multivariate Data Analysis with Readings*, 4th edn (Englewood Cliffs, NJ: Prentice Hall, 1995), 420–83.

11 For further discussion of the issues involved in standardisation, see Romsburg, H.C., *Cluster Analysis for Researchers* (Melbourne: Krieger Publishing, 1990).

12 Johnson, R.A. and Wichern, D.A., *Applied Multivariate Statistical Analysis*, 4th edn (Upper Saddle River, NJ: Prentice Hall, 1998); Milligan, G., 'An examination of the effect of six types of error perturbation on fifteen clustering algorithms', *Psychometrika* 45 (September 1980), 325–42.

13 Everitt, B., *Cluster Analysis*, 3rd edn (New York: Halstead Press, 1993); Punj, G. and Stewart, D., 'Cluster analysis in marketing research: reviews and suggestions for application', *Journal of Marketing Research* 20 (May 1983), 134–48.

14 For a formal discussion of reliability, validity and significance testing in cluster analysis, see Dibbs, S. and Stern, P., 'Questioning the reliability of market segmentation techniques', *Omega* 23(6) (December 1995), 625–36; Funkhouser, G.R., 'A note on the reliability of certain clustering algorithms', *Journal of Marketing Research* 30 (February 1983), 99–102; Klastorin, T.D., 'Assessing cluster analysis results', *Journal of Marketing Research* 20 (February 1983), 92–8; and Arnold, S.J., 'A test for clusters', *Journal of Marketing Research* 16 (November 1979), 545–51.

15 Saunders, J., Wong, V. and Doyle, P., 'The congruence of successful international competitors: a study of successful international competitors: a study of the marketing strategies and organisations of Japanese and US competitors in the UK', *Journal of Global Marketing* 7(3) (1994), 41–59; Doyle, P., Saunders, J. and Wong, V., 'International marketing strategies and organisations: a study of U.S., Japanese, and British competitors', in Bloom, P., Winer, R., Kassarjian, H.H., Scammon, D.L., Weitz, B., Spekman, R.E., Mahajan, V. and Levy, M. (eds), *Enhancing Knowledge Development in Marketing*, Series no. 55 (Chicago, IL: American Marketing Association, 1989), 100–4.

16 Holohean Jr, E.J., Banks, S.M. and Maddy, B.A., 'System impact and methodological issues in the development of an empirical typology of psychiatric hospital residents', *Journal of Mental Health Administration* 22(2) (Spring 1995), 177–88; Woodside, A.G., Nielsen, R.L., Walters, F. and Muller, G.D., 'Preference segmentation of health care services: the old-fashioneds, value conscious, affluents, and professional want-it-alls', *Journal of Health Care Marketing* (June 1988), 14–24.

17 Douglas, V., 'Questionnaire too long? Try variable clustering', *Marketing News* 29(5) (27 February 1995), 38.

18 Helgesen, T., 'The power of advertising – myths and realities', *Marketing and Research Today* 24(2) (May 1996), 63–71; Aaker, D.A., Stayman, D.M. and Vezina, R., 'Identifying feelings elicited by advertising', *Psychology and Marketing* (Spring 1988), 1–16.

19 Einspruch, E.L., *An Introductory Guide to SPSS for Windows* (Thousand Oaks, CA: Sage, 1998); Spector, P.E., *SAS Programming for Researchers and Social Scientists* (Thousand Oaks, CA: Sage, 1993); Norat, M.A., Software reviews, *Economic Journal: The Journal of the Royal Economic Society* 107 (May 1997), 857–82; Seiter, C., 'The statistical difference', *Macworld* 10(10) (October 1993), 116–21.