Cluster Analysis

9

Keywords

Agglomerative clustering • Average linkage • Canberra distance • Centroid linkage • Chaining effect • Chebychev distance • City-block distance • Clusters • Clustering variables • Complete linkage • Dendrogram • Distance matrix • Divisive clustering • Duda-Hart index • Euclidean distance • Factor-cluster segmentation • Gower's dissimilarity coefficient • Hierarchical clustering methods • Partitioning methods • k-means • k-medians • k-means++ • k-medoids • Label switching • Linkage algorithm • Local optimum • Mahalanobis distance • Manhattan metric • Market segmentation • Matching coefficients • Non-hierarchical clustering methods • Profiling • Russel and Rao coefficient • Single linkage • Simple matching coefficient • Straight line distance • Ties • Variance ration criterion • Ward's linkage • Weighted average linkage

Learning Objectives

After reading this chapter, you should understand:

- The basic concepts of cluster analysis.
- How basic cluster algorithms work.
- How to compute simple clustering results manually.
- The different types of clustering procedures.
- The Stata clustering outputs.

E. Mooi et al., *Market Research*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-5218-7_9

9.1 Introduction

Market segmentation is one of the most fundamental marketing activities. Since consumers, customers, and clients have different needs, companies have to divide markets into groups (segments) of consumers, customers, and clients with similar needs and wants. Firms can then target each of these segments by positioning themselves in a unique segment (e.g., Ferrari in the high-end sports car market). Market segmentation "is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien and Rangaswamy 2004, p. 61) and "tools such as segmentation [...] have the largest impact on marketing decisions" (John et al. 2014, p. 127). While market researchers often form market segments based on practical grounds, industry practice and wisdom, cluster analysis uses data to form segments, making segmentation less dependent on subjectivity.

9.2 Understanding Cluster Analysis

Cluster analysis is a method for segmentation and identifies homogenous groups of objects (or cases, observations) called **clusters**. These objects can be individual customers, groups of customers, companies, or entire countries. Objects in a certain cluster should be as similar as possible to each other, but as distinct as possible from objects in other clusters.

Let's try to gain a basic understanding of cluster analysis by looking at a simple example. Imagine that you are interested in segmenting your customer base in order to better target them through, for example, pricing strategies.

The first step is to decide on the characteristics that you will use to segment your customers A to G. In other words, you have to decide which **clustering variables** will be included in the analysis. For example, you may want to segment a market based on customers' price consciousness (x) and brand loyalty (y). These two variables can be measured on a scale from 0 to 100 with higher values denoting a higher degree of price consciousness and brand loyalty. Table 9.1 and the scatter plot in Fig. 9.1 show the values of seven customers (referred to as objects).

The aim of cluster analysis is to identify groups of objects (in this case, customers) that are very similar regarding their price consciousness and brand loyalty, and assign them to clusters. After having decided on the clustering variables (here, price consciousness and brand loyalty), we need to decide on the clustering procedure to form our groups of objects. This step is crucial for the analysis, as different procedures require different decisions prior to analysis. There is an abundance of different approaches and little guidance on which one to use in practice. We will discuss the most popular approaches in market research, including:

- hierarchical methods, and
- partitioning methods (more precisely k-means)

Table 9.1 Data

Customer	A	B	C	D	E	F	G
x	33	82	66	30	79	50	10
y	95	94	80	67	60	33	17

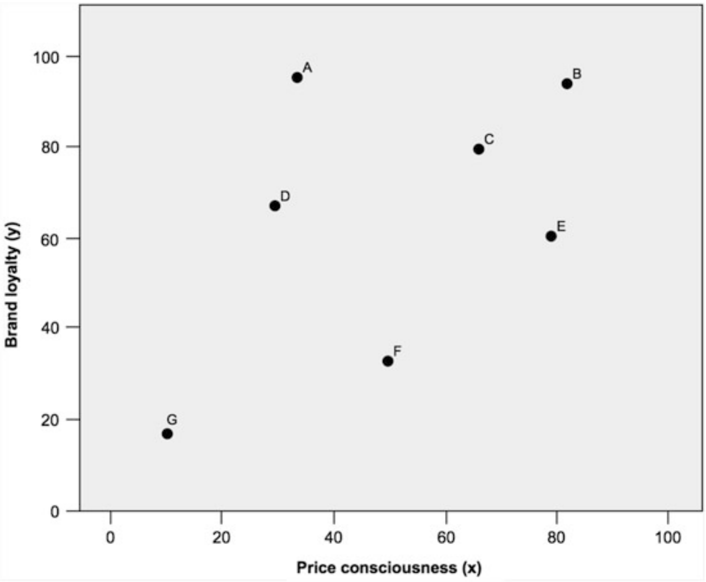


Fig. 9.1 Scatter plot

While the basic aim of these procedures is the same, namely grouping similar objects into clusters, they take different routes, which we will discuss in this chapter. An important consideration before starting the grouping is to determine how similarity should be measured. Most methods calculate measures of (dis) similarity by estimating the distance between pairs of objects. Objects with smaller distances between one another are considered more similar, whereas objects with larger distances are considered more dissimilar. The decision on how many clusters should be derived from the data is a fundamental issue in the application of cluster analysis. This question is explored in the next step of the analysis. In most instances, we do not know the exact number of clusters and then we face a trade-off. On the one hand, we want as few clusters as possible to make the clusters easy to understand and actionable. On the other hand, having many clusters allows us to identify subtle differences between objects.

Megabus is a hugely successful bus line in the US. They completely rethought the nature of their customers and concentrated on three specific segments of the market: College kids, women travelling in groups, and active seniors. To meet these customer segments' needs, Megabus reimagined the entire driving experience by developing double-decker buses with glass roofs and big windows, and equipped with fast WiFi. In light of the success of Megabus's segmenting and targeting efforts, practitioners even talk about the "Megabus Effect"—how one company has shaped an entire industry.

In the final step, we need to interpret the clustering solution by defining and labeling the obtained clusters. We can do so by comparing the mean values of the clustering variables across the different clusters, or by identifying explanatory variables to profile the clusters. Ultimately, managers should be able to identify customers in each cluster on the basis of easily measurable variables. This final step also requires us to assess the clustering solution's stability and validity. Figure 9.2 illustrates the steps associated with a cluster analysis; we will discuss these steps in more detail in the following sections.

9.3 Conducting a Cluster Analysis

9.3.1 Select the Clustering Variables

At the beginning of the clustering process, we have to select appropriate variables for clustering. Even though this choice is critical, it is rarely treated as such. Instead, a mixture of intuition and data availability guide most analyses in marketing

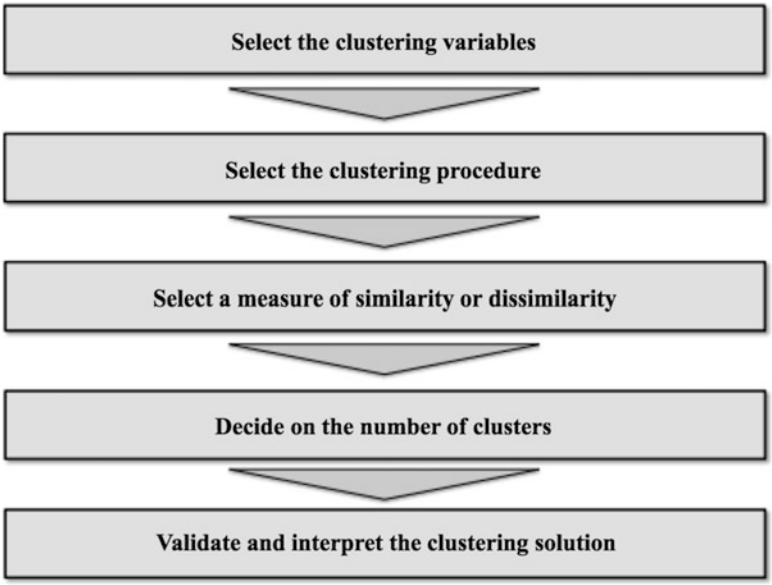

Select the clustering variables
Select the clustering procedure
Select a measure of similarity or dissimilarity
Decide on the number of clusters
Validate and interpret the clustering solution

Fig. 9.2 Steps in a cluster analysis

practice. However, faulty assumptions may lead to improper market segmentation and, consequently, to deficient marketing strategies. Thus, great care should be taken when selecting the clustering variables! There are several types of clustering variables, as shown in Fig. 9.3. Sociodemographic variables define clusters based on people's demographic (e.g., age, ethnicity, and gender), geographic (e.g., residence in terms of country, state, and city), and socioeconomic (e.g., education, income, and social class) characteristics. Psychometric variables capture unobservable character traits such as people's personalities or lifestyles. Finally, behavioral clustering variables typically consider different facets of consumer behavior, such as the way people purchase, use, and dispose of products. Other behavioral clustering variables capture specific benefits which different groups of consumers look for in a product.

The types of variables used for cluster analysis provide different solutions and, thereby, influence targeting strategies. Over the last decades, attention has shifted from more traditional sociodemographic clustering variables towards behavioral and psychometric variables. The latter generally provide better guidance for decisions on marketing instruments' effective specification. Generally, clusters based on psychometric variables are more homogenous and these consumers respond more consistently to marketing actions (e.g., Wedel and Kamakura 2000). However, consumers in these clusters are frequently hard to identify as such variables are not easily measured. Conversely, clusters determined by sociodemographic variables are easy to identify but are also more heterogeneous, which complicates targeting efforts. Consequently, researchers frequently combine

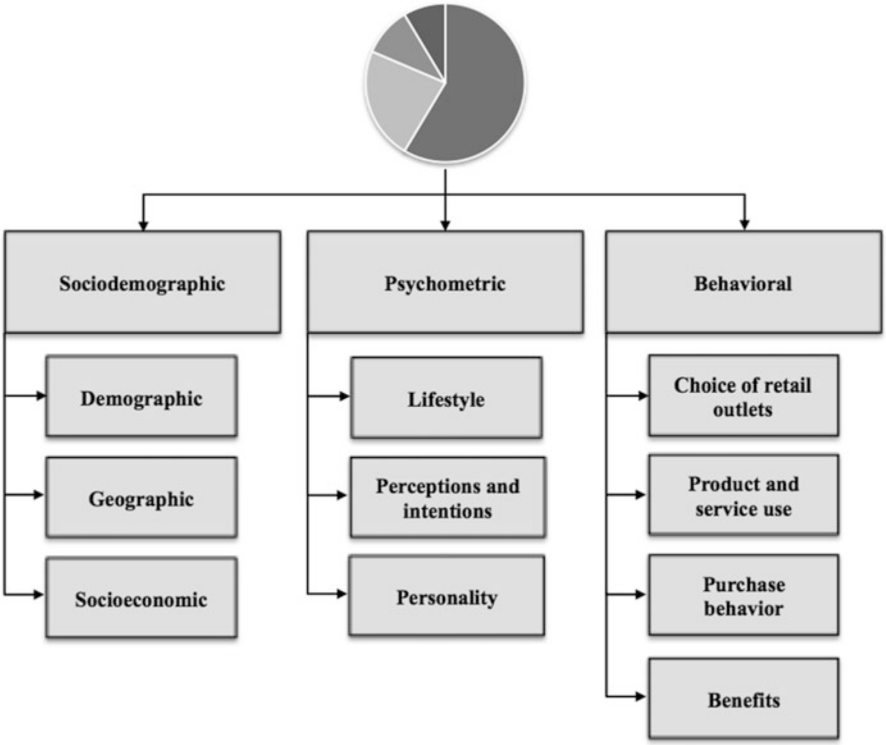


Fig. 9.3 Types of clustering variables

different variables such as lifestyle characteristics and demographic variables, benefiting from each one's strengths.

In some cases, the choice of clustering variables is apparent because of the task at hand. For example, a managerial problem regarding corporate communications will have a fairly well defined set of clustering variables, including contenders such as awareness, attitudes, perceptions, and media habits. However, this is not always the case and researchers have to choose from a set of candidate variables. But how do we make this decision? To facilitate the choice of clustering variables, we should consider the following guiding questions:

– Do the variables differentiate sufficiently between the clusters?
– Is the relation between the sample size and the number of clustering variables reasonable?
– Are the clustering variables highly correlated?
– Are the data underlying the clustering variables of high quality?

Do the variables differentiate sufficiently between the clusters?

It is important to select those clustering variables that provide a clear-cut differentiation between the objects.[1] More precisely, criterion validity is of special interest; that is, the extent to which the "independent" clustering variables are associated with one or more criterion variables not included in the analysis. Such criterion variables generally relate to an aspect of behavior, such as purchase intention or willingness-to-pay. Given this relationship, there should be significant differences between the criterion variable(s) across the clusters (e.g., consumers in one cluster exhibit a significantly higher willingness-to-pay than those in other clusters). These associations may or may not be causal, but it is essential that the clustering variables distinguish significantly between the variable(s) of interest.

Is the relation between the sample size and the number of clustering variables reasonable?

When choosing clustering variables, the sample size is a point of concern. First and foremost, this relates to issues of managerial relevance as the cluster sizes need to be substantial to ensure that the targeted marketing programs are profitable. From a statistical perspective, every additional variable requires an over-proportional increase in observations to ensure valid results. Unfortunately, there is no generally accepted guideline regarding minimum sample sizes or the relationship between the objects and the number of clustering variables used. While early research suggested a minimum sample size of two to the power of the number of clustering variables (Formann 1984), more recent rules-of-thumb are as follows:

- In the simplest case where clusters are of equal size, Qiu and Joe (2009) recommend a sample size at least ten times the number of clustering variables multiplied by the number of clusters.
- Dolnicar et al. (2014) recommend using a sample size of 70 times the number of clustering variables.
- Dolnicar et al. (2016) find that increasing the sample size from 10 to 30 times the number of clustering variables substantially improves the clustering solution. This improvement levels off subsequently, but is still noticeable up to a sample size of approximately 100 times the number of clustering variables.

These rules-of-thumb provide only rough guidance as the required sample size depends on many factors, such as the survey data characteristics (e.g., nonresponse, sampling error, response styles), relative cluster sizes, and the degree to which the clusters overlap (Dolnicar et al. 2016). However, these rules also jointly suggest that a minimum of 10 times the number of clustering variables should be considered the bare minimum. Keep in mind that no matter how many variables are used and

[1] Tonks (2009) provides a discussion of segment design and the choice of clustering variables in consumer markets.

no matter how small the sample size, cluster analysis will almost always provide a result. At the same time, however, the quality of results shows decreasing marginal returns as the sample size increases. Since cluster analysis is an exploratory technique whose results should be interpreted by taking practical considerations into account, it is not necessary to increase the sample size massively.

Are the clustering variables highly correlated?

If there is strong correlation between the variables, they are not sufficiently unique to identify distinct market segments. If highly correlated variables are used for cluster analysis, the specific aspects that these variables cover will be overrepresented in the clustering solution. In this regard, absolute correlations above 0.90 are always problematic. For example, if we were to add another variable called *brand preference* to our analysis, it would almost cover the same aspect as *brand loyalty*. The concept of being attached to a brand would therefore be overrepresented in the analysis, because the clustering procedure does not conceptually differentiate between the clustering variables. Researchers frequently handle such correlation problems by applying cluster analysis to the observations' factor scores derived from a previously carried out principal component or factor analysis. However, this **factor-cluster segmentation** approach is subject to several limitations, which we discuss in Box 9.1.

Box 9.1 Issues with Factor-Cluster Segmentation

Dolnicar and Grün (2009) identify several problems of the factor-cluster segmentation approach (see Chap. 8 for a discussion of principal component and factor analysis and related terminology):

1. The data are pre-processed and the clusters are identified on the basis of transformed values, not on the original information, which leads to different results.
2. In factor analysis, the factor solution does not explain all the variance; information is thus discarded before the clusters have been identified or constructed.
3. Eliminating variables with low loadings on all the extracted factors means that, potentially, the most important pieces of information for the identification of niche clusters are discarded, making it impossible to ever identify such groups.
4. The interpretations of clusters based on the original variables become questionable, given that these clusters were constructed by using factor scores.

(continued)

Box 9.1 (continued)

Several studies have shown that the factor-cluster segmentation reduces the success of finding useable clusters significantly.[2] Consequently, you should reduce the number of items in the questionnaire's pre-testing phase, retaining a reasonable number of relevant, non-overlapping questions that you believe differentiate the clusters well. However, if you have doubts about the data structure, factor-clustering segmentation may still be a better option than discarding items.

Are the data underlying the clustering variables of high quality?

Ultimately, the choice of clustering variables always depends on contextual influences, such as the data availability or the resources to acquire additional data. Market researchers often overlook that the choice of clustering variables is closely connected to data quality. Only those variables that ensure that high quality data can be used should be included in the analysis (Dolnicar and Lazarevski 2009). Following our discussions in Chaps. 3, 4 and 5, data are of high quality if the questions. . .

– . . . have a strong theoretical basis,
– . . . are not contaminated by respondent fatigue or response styles, and
– . . . reflect the current market situation (i.e., they are recent).

The requirements of other functions in the organization often play a major role in the choice of clustering variables. Consequently, you have to be aware that the choice of clustering variables should lead to segments acceptable to the different functions in the organization.

9.3.2 Select the Clustering Procedure

By choosing a specific clustering procedure, we determine how clusters should be formed. This forming of clusters always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables' overall variance of the objects in a specific cluster), or maximizing the distance between the clusters. The procedure could also address the question of how to determine the (dis)similarity between objects in a newly formed cluster and the remaining objects in the dataset.

There are many different clustering procedures and also many ways of classifying these (e.g., overlapping versus non-overlapping, unimodal versus

[2] See Arabie and Hubert (1994), Sheppard (1996), and Dolnicar and Grün (2009).

multimodal, exhaustive versus non-exhaustive). Wedel and Kamakura (2000), Dolnicar (2003), and Kaufman and Rousseeuw (2005) offer reviews of clustering techniques. A practical distinction is the differentiation between hierarchical and partitioning methods (especially *k*-means), which we will discuss in the next sections.

9.3.2.1 Hierarchical Clustering Methods

Understanding Hierarchical Clustering Methods

Hierarchical clustering methods are characterized by the tree-like structure established in the course of the analysis. Most hierarchical methods fall into a category called **agglomerative clustering**. In this category, clusters are consecutively formed from objects. Agglomerative clustering starts with each object representing an individual cluster. The objects are then sequentially merged to form clusters of multiple objects, starting with the two most similar objects. Similarity is typically defined in terms of the distance between objects. That is, objects with smaller distances between one another are considered more similar, whereas objects with larger distances are considered more dissimilar. After the merger of the first two most similar (i.e., closest) objects, the agglomerative clustering procedure continues by merging another pair of objects or adding another object to an already existing cluster. This procedure continues until all the objects have been merged into one big cluster. As such, agglomerative clustering establishes a hierarchy of objects from the bottom (where each object represents a distinct cluster) to the top (where all objects form one big cluster). The left-hand side of Fig. 9.4 shows how agglomerative clustering merges objects (represented by circles) step-by-step with other objects or clusters (represented by ovals).

Hierarchical clustering can also be interpreted as a top-down process, where all objects are initially merged into a single cluster, which the algorithm then gradually splits up. This approach to hierarchical clustering is called **divisive clustering**. The right-hand side of Fig. 9.4 illustrates the divisive clustering concept. As we can see, in both agglomerative and divisive clustering, a cluster on a higher level of the hierarchy always encompasses all clusters from a lower level. This means that if an object is assigned to a certain cluster, there is no possibility of reassigning this object to another cluster (hence, hierarchical clustering). This is an important distinction between hierarchical and partitioning methods, such as *k*-means, which we will explore later in this chapter.

Divisive procedures are rarely used in market research and not implemented in statistical software programs such as Stata as they are computationally very intensive for all but small datasets.[3] We therefore focus on (agglomerative) hierarchical clustering.

[3] Whereas agglomerative methods have the large task of checking $N \cdot (N-1)/2$ possible first combinations of observations (note that N represents the number of observations in the dataset), divisive methods have the almost impossible task of checking $2^{(N-1)}-1$ combinations.

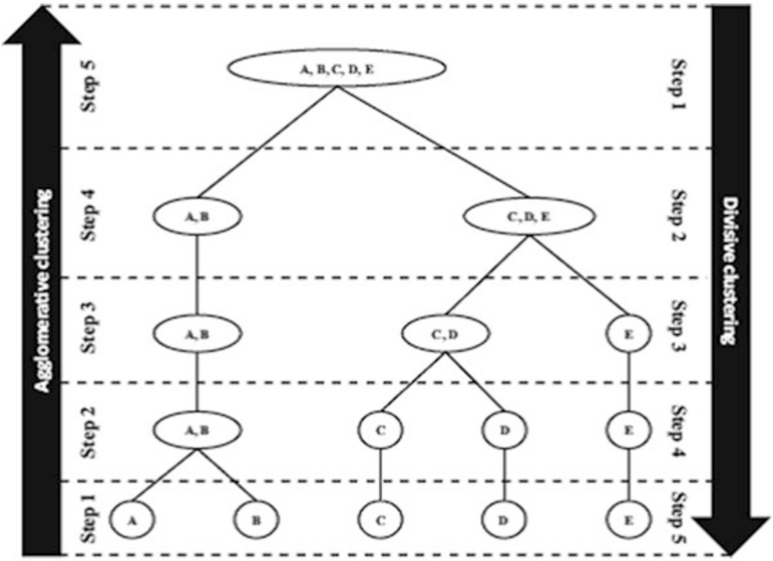


Fig. 9.4 Agglomerative and divisive clustering

Linkage algorithms

When using agglomerative hierarchical clustering, you need to specify a **linkage algorithm.** Linkage algorithms define the distance from a newly formed cluster to a certain object, or to other clusters in the solution. The most popular linkage algorithms include the following:

- **Single linkage** (nearest neighbor): The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.
- **Complete linkage** (furthest neighbor): The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.
- **Average linkage**: The distance between two clusters is defined as the average distance between all pairs of the two clusters' members. **Weighted average linkage** performs the same calculation, but weights distances based on the number of objects in the cluster. Thus, the latter method is preferred when clusters are not of approximately equal size.
- **Centroid linkage**: In this approach, the geometric center (centroid) of each cluster is computed first. This is done by computing the clustering variables' average values of all the objects in a certain cluster. The distance between the two clusters equals the distance between the two centroids.
- **Ward's linkage**: This approach differs from the previous ones in that it does not combine the two closest or most similar objects successively. Instead, Ward's linkage combines those objects whose merger increases the overall within-

cluster variance (i.e., the homogeneity of clusters) to the smallest possible degree. The approach is generally used in combination with (squared) Euclidean distances, but can be used in combination with any other (dis)similarity measure.

Figures 9.5, 9.6, 9.7, 9.8 and 9.9 illustrate these linkage algorithms for two clusters, which are represented by white circles surrounding a set of objects. Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has specific properties:

- The single linkage algorithm is based on minimum distances; it tends to form one large cluster with the other clusters containing only one or a few objects each. We can make use of this **chaining effect** to detect outliers, as these will be merged with the remaining objects—usually at very large distances—in the last steps of the analysis. Single linkage is considered the most versatile algorithm.
- The complete linkage method is strongly affected by outliers, as it is based on maximum distances. Clusters produced by this method are likely to be compact and tightly clustered.
- The average linkage and centroid linkage algorithms tend to produce clusters with low within-cluster variance and with similar sizes. The average linkage is affected by outliers, but less than the complete linkage method.
- Ward's linkage yields clusters of similar size with a similar degree of tightness. Prior research has shown that the approach generally performs very well. However, outliers and highly correlated variables have a strong bearing on the algorithm.

To better understand how the linkage algorithms work, let's manually examine some calculation steps using single linkage as an example. Let's start by looking at the distance matrix in Table 9.2, which shows the distances between objects A-G from our initial example. In this distance matrix, the non-diagonal elements express the distances between pairs of objects based on the Euclidean distance—we will discuss this distance measure in the following section. The diagonal elements of the matrix represent the distance from each object to itself, which is, of course, 0. In our example, the distance matrix is an 8 × 8 table with the lines and rows representing the objects under consideration (see Table 9.1). As the distance between objects B and C (in this case, 21.260 units) is the same as between C and B, the distance matrix is symmetrical. Furthermore, since the distance between an object and itself is 0, you only need to look at either the lower or upper non-diagonal elements.

In the very first step, the two objects exhibiting the smallest distance in the matrix are merged. Since the smallest distance occurs between B and C ($d(B,C) = 21.260$; printed in bold in Table 9.2), we merge these two objects in the first step of the analysis.

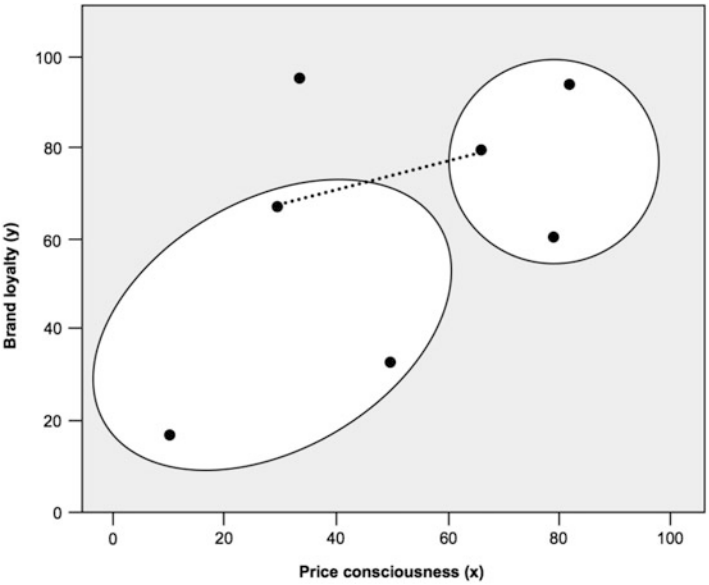


Fig. 9.5 Single linkage

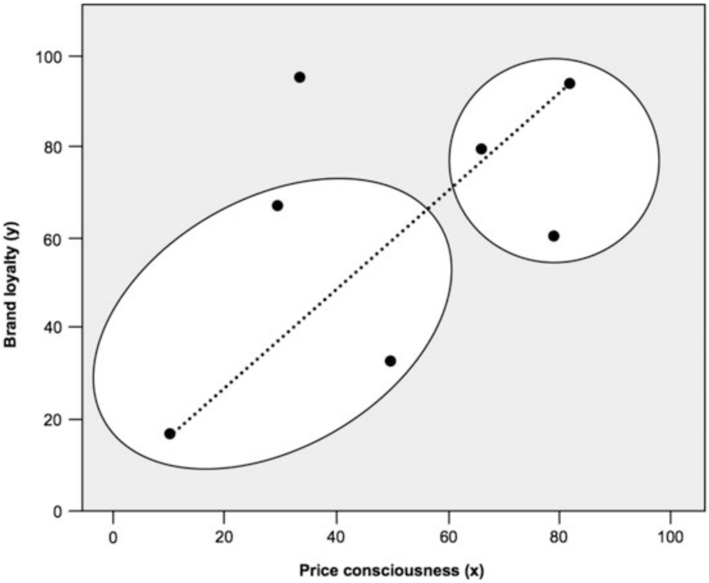


Fig. 9.6 Complete linkage

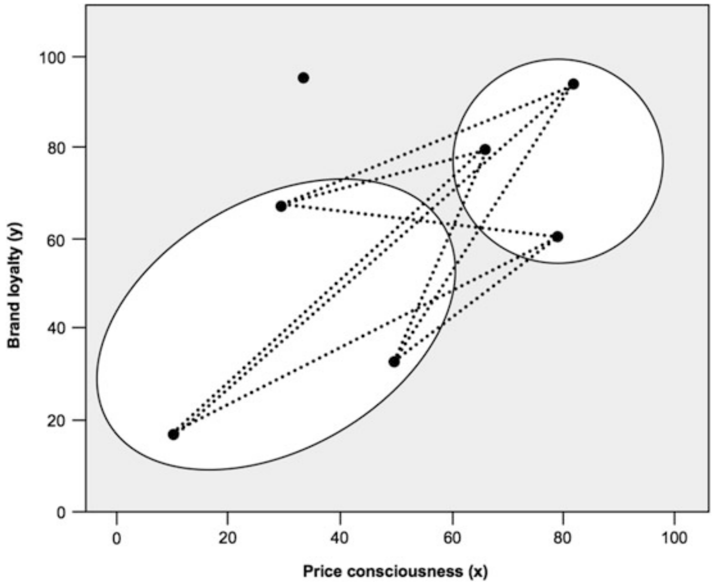


Fig. 9.7 Average linkage

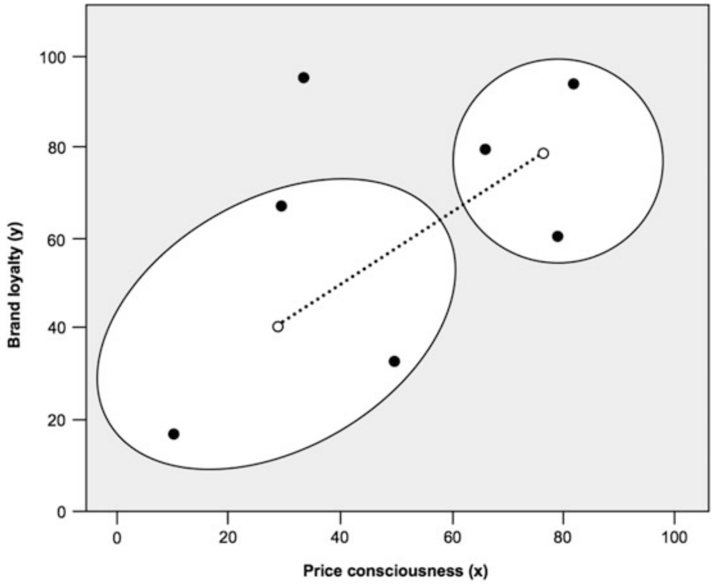


Fig. 9.8 Centroid linkage

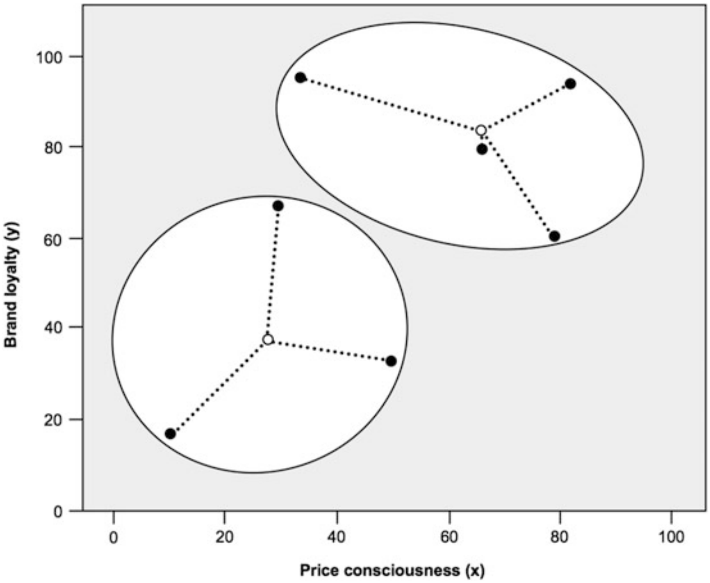


Fig. 9.9 Ward's linkage

Agglomerative clustering procedures always merge those objects with the smallest distance, regardless of the linkage algorithm used (e.g., single or complete linkage).

In the next step, we form a new distance matrix by considering the single linkage decision rule as discussed above. Using this linkage algorithm, we need to compute the distance from the newly formed cluster [B,C] (clusters are indicated by squared brackets) to all the other objects. For example, with regard to the distance from the cluster [B,C] to object A, we need to check whether A is closer to object B or to object C. That is, we look for the minimum value in d(A,B) and d(A,C) from Table 9.2. As d(A,C) = 36.249 is smaller than d(A,B) = 49.010, the distance from A to the newly formed cluster is equal to d(A,C); that is, 36.249. We also compute the distances from cluster [B,C] to all the other objects (i.e., D, E, F, G). For example, the distance between [B,C] and D is the minimum of d(B,D) = 58.592 and d(C,D) = 38.275 (Table 9.2). Finally, there are several distances, such as d(D,E) and d(E,F), which are not affected by the merger of B and C. These distances are simply copied into the new distance matrix. This yields the new distance matrix shown in Table 9.3.

Continuing the clustering procedure, we simply repeat the last step by merging the objects in the new distance matrix that exhibit the smallest distance and calculate the distance from this new cluster to all the other objects. In our case,

Table 9.2 Euclidean distance matrix

Objects	A	B	C	D	E	F	G
A	0						
B	49.010	0					
C	36.249	**21.260**	0				
D	28.160	58.592	38.275	0			
E	57.801	34.132	23.854	40.497	0		
F	64.288	68.884	49.649	39.446	39.623	0	
G	81.320	105.418	84.291	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold

Table 9.3 Distance matrix after first clustering step (single linkage)

Objects	A	B, C	D	E	F	G
A	0					
B, C	36.249	0				
D	28.160	38.275	0			
E	57.801	**23.854**	40.497	0		
F	64.288	49.649	39.446	39.623	0	
G	81.320	84.291	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold

Table 9.4 Distance matrix after second clustering step (single linkage)

Objects	A	B, C, E	D	F	G
A	0				
B, C, E	36.249	0			
D	**28.160**	38.275	0		
F	64.288	39.623	39.446	0	
G	81.320	81.302	53.852	43.081	0

Note: Smallest distance is printed in bold

Table 9.5 Distance matrix after third clustering step (single linkage)

Objects	A, D	B, C, E	F	G
A, D	0			
B, C, E	**36.249**	0		
F	39.446	39.623	0	
G	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold

the smallest distance (23.854, printed in bold in Table 9.3) occurs between the newly formed cluster [B, C] and object E. The result of this step is described in Table 9.4.

Try to calculate the remaining steps yourself and compare your solution with the distance matrices in the following Tables 9.5, 9.6 and 9.7.

Table 9.6 Distance matrix after fourth clustering step (single linkage)

Objects	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	**39.446**	0	
G	53.852	43.081	0

Note: Smallest distance is printed in bold

Table 9.7 Distance matrix after fifth clustering step (single linkage)

Objects	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	43.081	0

By following the single linkage procedure, the last steps involve the merger of cluster [A,B,C,D,E,F] and object G at a distance of 43.081. Do you get the same results? As you can see, conducting a basic cluster analysis manually is not that hard at all—not if there are only a few objects.

9.3.2.2 Partitioning Methods: *k*-means

Partitioning clustering methods are another important group of procedures. As with hierarchical clustering, there is a wide array of different algorithms; of these, *k*-means is the most popular for market research.

Understanding *k*-means Clustering

The ***k*-means** method follows an entirely different concept than the hierarchical methods discussed above. The initialization of the analysis is one crucial difference. Unlike with hierarchical clustering, we need to specify the number of clusters to extract from the data prior to the analysis. Using this information as input, *k*-means then assigns all the objects to the number of clusters that the researcher specifies. This starting partition comes in different forms. Examples of these forms include:

- randomly select *k* objects as starting centers for the *k* clusters (*K unique random observations* in Stata),
- use the first or last *k* objects as starting centers for the *k* clusters (*First K observations* and *Last K observations* in Stata),
- randomly allocate all the objects into *k* groups and compute the means (or medians) of each group. These means (or medians) then serve as starting centers (*Group means from K random partitions of the data* in Stata), and
- provide an initial grouping variable that defines the groups among the objects to be clustered. The group means (or medians) of these groups are used as the starting centers (*Group means from partitions defined by initial grouping variables* in Stata).

After the initialization, *k*-means successively reassigns the objects to other clusters with the aim of minimizing the within-cluster variation. This within-cluster variation is equal to the squared distance of each observation to the center of the associated cluster (i.e., the centroid). If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster.

Since cluster affiliations can change in the course of the clustering process (i.e., an object can move to another cluster in the course of the analysis), *k*-means does not build a hierarchy, which hierarchical clustering does (Fig. 9.4). Therefore, *k*-means belongs to the group of **non-hierarchical clustering methods**.

For a better understanding of the approach, let's take a look at how it works in practice. Figures 9.10, 9.11, 9.12 and 9.13 illustrate the four steps of the *k*-means clustering process—research has produced several variants of the original algorithm, which we briefly discuss in Box 9.2.

- **Step 1:** The researcher needs to specify the number of clusters that *k*-means should retain from the data. Using this number as the input, the algorithm selects a center for each cluster. In our example, two cluster centers are randomly initiated, which CC1 (first cluster) and CC2 (second cluster) represent in Fig. 9.10.
- **Step 2:** Euclidean distances are computed from the cluster centers to every object. Each object is then assigned to the cluster center with the shortest distance to it. In our example (Fig. 9.11), objects A, B, and C are assigned to the first cluster, whereas objects D, E, F, and G are assigned to the second. We now have our initial partitioning of the objects into two clusters.
- **Step 3:** Based on the initial partition in step 2, each cluster's geometric center (i.e., its centroid) is computed. This is done by computing the mean values of the objects contained in the cluster (e.g., A, B, C in the first cluster) in terms of each of the variables (price consciousness and brand loyalty). As we can see in Fig. 9.12, both clusters' centers now shift to new positions (CC1' in the first and CC2' in the second cluster; the inverted comma indicates that the cluster center has changed).
- **Step 4:** The distances are computed from each object to the newly located cluster centers and the objects are again assigned to a certain cluster on the basis of their minimum distance to other cluster centers (CC1' and CC2'). Since the cluster centers' position changed with respect to the initial situation in the first step, this could lead to a different cluster solution. This is also true of our example, because object E is now—unlike in the initial partition—closer to the first cluster center (CC1') than to the second (CC2'). Consequently, this object is now assigned to the first cluster (Fig. 9.13).

The *k*-means procedure is now repeated until a predetermined number of iterations are reached, or convergence is achieved (i.e., there is no change in the cluster affiliations).

Three aspects are worth noting in terms of using *k*-means:

- *k*-means is implicitly based on pairwise Euclidean distances, because the sum of the squared distances from the centroid is equal to the sum of the pairwise squared Euclidean distances divided by the number of objects. Therefore, the method should only be used with metric and, in case of equidistant scales, ordinal variables. Similarly, you should only use (squared) Euclidean distances with *k*-means.

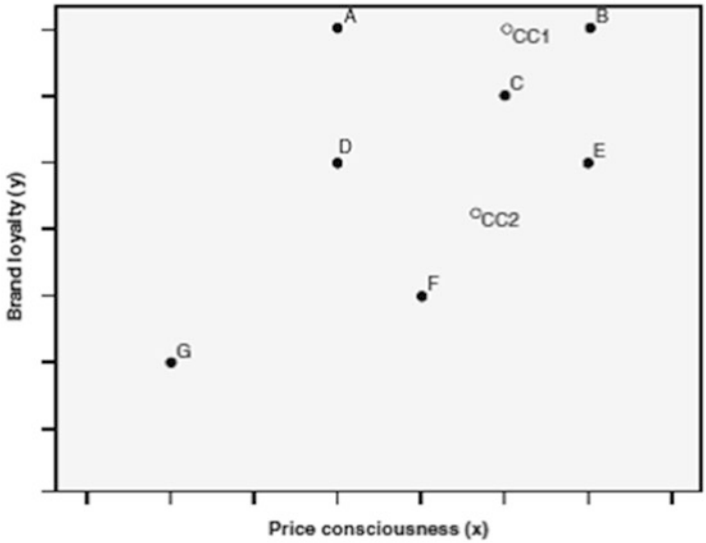


Fig. 9.10 *k*-means procedure (step 1: placing random cluster centers)

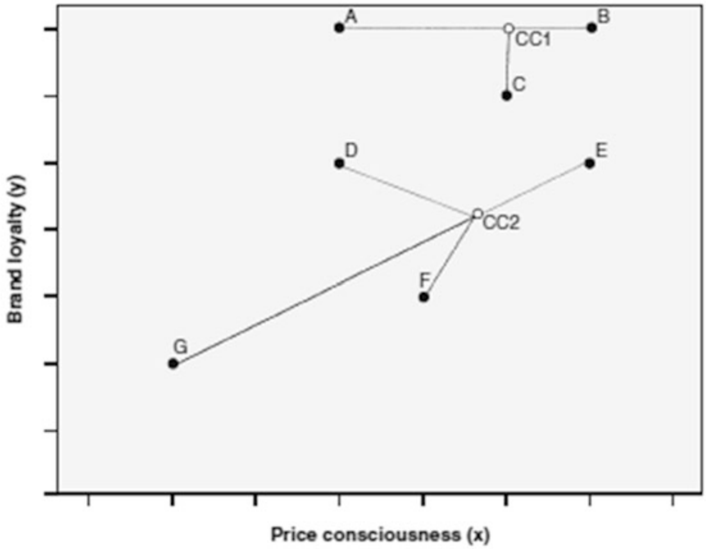


Fig. 9.11 *k*-means procedure (step 2: assigning objects to the closest cluster center)

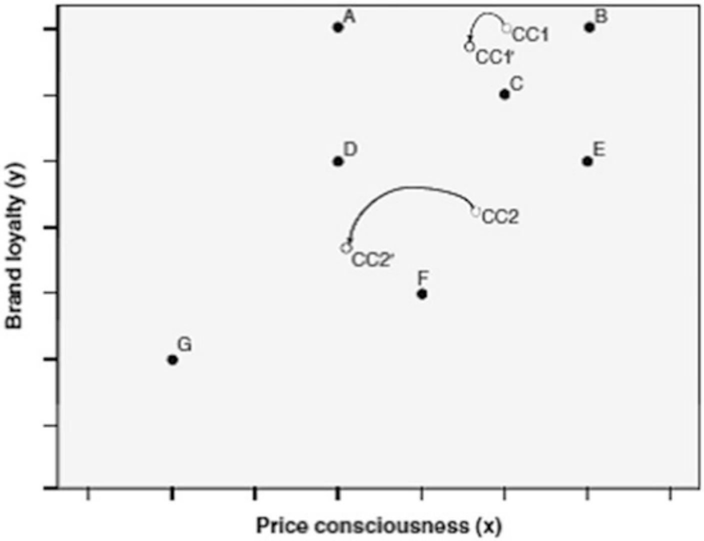


Fig. 9.12 *k*-means procedure (step 3: recomputing cluster centers)

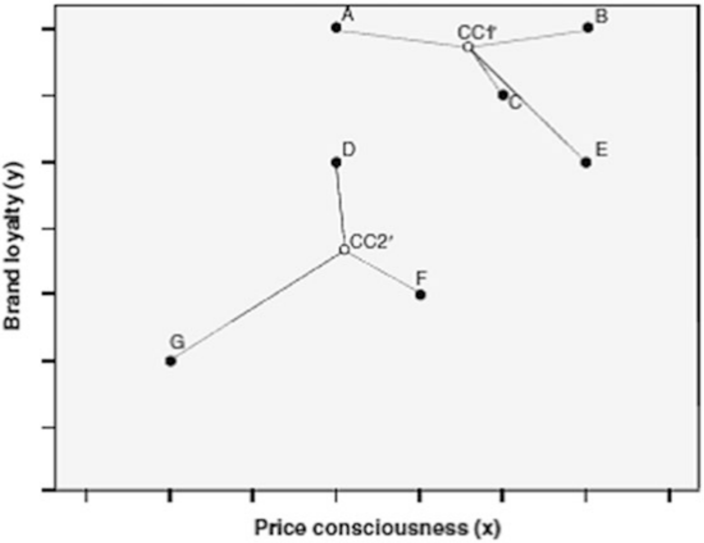


Fig. 9.13 *k*-means procedure (step 4: reassigning objects to the closest cluster center)

– Results produced by *k*-means depend on the starting partition. That is, *k*-means produce different results, depending on the starting partition chosen by the researcher or randomly initiated by the software. As a result, *k*-means may converge in a **local optimum**, which means that the solution is only optimal compared to similar solutions, but not globally. Therefore, you should run *k*-means multiple times using different options for generating a starting partition.

- *k*-means is less computationally demanding than hierarchical clustering techniques. The method is therefore generally preferred for sample sizes above 500, and particularly for *big data* applications.
- Running *k*-means requires specifying the number of clusters to retain prior to running the analysis. We discuss this issue in the next section.

Box 9.2 Variants of the Original *k*-means Method

k*-medians** is a popular variant of *k*-means and has also been implemented in Stata. This procedure essentially follows the same logic and procedure as *k*-means. However, instead of using the cluster mean as a reference point for the calculation of the within cluster variance, *k*-medians minimizes the absolute deviations from the cluster medians, which equals the city-block distance. Thus, *k*-medians does *not* optimize the squared deviations from the mean as in *k*-means, but absolute distances. In this way, *k*-medians avoids the possible effect of extreme values on the cluster solution. Further variants, which are not menu-accessible in Stata, use other cluster centers (e.g., ***k-medoids; Kaufman and Rousseeuw 2005; Park and Jun 2009), or optimize the initialization process (e.g., ***k-means++***; Arthur and Vassilvitskii 2007).

9.3.3 Select a Measure of Similarity or Dissimilarity

In the previous section, we discussed different linkage algorithms used in agglomerative hierarchical clustering as well as the *k*-means procedure. All these clustering procedures rely on measures that express the (dis)similarity between pairs of objects. In the following section, we introduce different measures for metric, ordinal, nominal, and binary variables.

9.3.3.1 Metric and Ordinal Variables

Distance Measures

A straightforward way to assess two objects' proximity is by drawing a straight line between them. For example, on examining the scatter plot in Fig. 9.1, we can easily see that the length of the line connecting observations B and C is much shorter than the line connecting B and G. This type of distance is called **Euclidean distance** or **straight line distance**; it is the most commonly used type for analyzing metric variables and, if the scales are equidistant (Chap. 3), ordinal variables. Statistical software programs such as Stata simply refer to the Euclidean distance as *L2*, as it is a specific type of the more general Minkowski distance metric with argument 2 (Anderberg 1973). Researchers also often use the **squared Euclidean distance**, referred to as *L2 squared* in Stata. For *k*-means, using the squared Euclidean distance is more appropriate because of the way the method computes the distances from the objects to the centroids (see Section 9.3.2.2).

In order to use a hierarchical clustering procedure, we need to express these distances mathematically. Using the data from Table 9.1, we can compute the Euclidean distance between customer B and customer C (generally referred to as d(B,C)) by using variables x and y with the following formula:

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

As can be seen, the Euclidean distance is the square root of the sum of the squared differences in the variables' values. Using the data from Table 9.1, we obtain the following:

$$d_{Euclidean}(B,C) = \sqrt{(82-66)^2 + (94-80)^2 =}\sqrt{452} \approx 21.260$$

This distance corresponds to the length of the line that connects objects B and C. In this case, we only used two variables, but we can easily add more under the root sign in the formula. However, each additional variable will add a dimension to our research problem (e.g., with six clustering variables, we have to deal with six dimensions), making it impossible to represent the solution graphically. Similarly, we can compute the distance between customer B and G, which yields the following:

$$d_{Euclidean}(B,G) = \sqrt{(82-10)^2 + (94-17)^2 =}\sqrt{11,113} \approx 105.418$$

Likewise, we can compute the distance between all other pairs of objects and summarize them in a distance matrix. Table 9.2, which we used as input to illustrate the single linkage algorithm, shows the Euclidean distance matrix for objects A-G.

There are also alternative distance measures: The **city-block distance** (called *L1* in Stata) uses the sum of the variables' absolute differences. This distance measure is referred to as the **Manhattan metric** as it is akin to the walking distance between two points in a city like New York's Manhattan district, where the distance equals the number of blocks in the directions North-South and East-West. Using the city-block distance to compute the distance between customers B and C (or C and B) yields the following:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |82-66| + |94-80| = 30$$

The resulting distance matrix is shown in Table 9.8.

Lastly, when working with metric (or ordinal) data, researchers frequently use the **Chebychev distance** (called *Linfinity* in Stata), which is the maximum of the

Table 9.8 City-block distance matrix

Objects	A	B	C	D	E	F	G
A	0						
B	50	0					
C	48	30	0				
D	31	79	49	0			
E	81	37	33	56	0		
F	79	93	63	54	56	0	
G	101	149	119	70	112	56	0

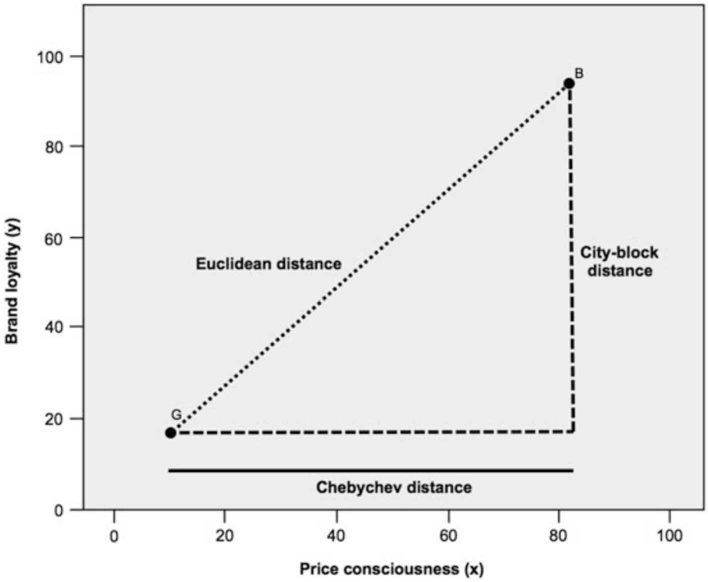


Fig. 9.14 Distance measures

absolute difference in the clustering variables' values. In respect of customers B and C, this result is:

$$d_{Chebychev}(B, C) = \max(|x_B - x_C|, |y_B - y_C|) = \max(|82 - 66|, |94 - 80|) = 16$$

Figure 9.14 illustrates the interrelation between these three distance measures regarding two objects (here: B and G) from our example.

Research has brought forward a range of other distance measures suitable for specific research settings. For example, the Stata menu offers the **Canberra distance**, a weighted version of the city-block distance, which is typically used for clustering data scattered widely around an origin. Other distance measures, such as the **Mahalanobis distance**, which compensates for collinearity between the clustering variables, are accessible via Stata syntax.

Different distance measures typically lead to different cluster solutions. Thus, it is advisable to use several measures, check for the stability of results, and compare them with theoretical or known patterns.

Association Measures

The (dis)similarity between objects can also be expressed by means of *association measures* (e.g., correlations). For example, suppose a respondent rated price consciousness 2 and brand loyalty 3, a second respondent indicated 5 and 6, whereas a third rated these variables 3 and 3. Euclidean and city-block, distances indicate that the first respondent is more similar to the third than to the second. Nevertheless, one could convincingly argue that the first respondent's ratings are more similar to the second's, as both rate brand loyalty higher than price consciousness. This can be accounted for by computing the correlation between two vectors of values as a measure of similarity (i.e., high correlation coefficients indicate a high degree of similarity). Consequently, similarity is no longer defined by means of the difference between the answer categories, but by means of the similarity of the answering profiles.

Whether you use one of the distance measures or correlations depends on whether you think the relative magnitude of the variables within an object (which favors correlation) matters more than the relative magnitude of each variable across the objects (which favors distance). Some researchers recommended using correlations when applying clustering procedures that are particularly susceptible to outliers, such as complete linkage, average linkage or centroid linkage. Furthermore, correlations implicitly standardize the data, as differences in the scale categories do not have a strong bearing on the interpretation of the response patterns. Nevertheless, distance measures are most commonly used for their intuitive interpretation. Distance measures best represent the concept of proximity, which is fundamental to cluster analysis. Correlations, although having widespread application in other techniques, represent patterns rather than proximity.

Standardizing the Data

In many analysis tasks, the variables under consideration are measured in different units with hugely different variance. This would be the case if we extended our set of clustering variables by adding another metric variable representing the customers' gross annual income. Since the absolute variation of the income variable would be much higher than the variation of the remaining two variables (remember, x and y are measured on a scale from 0 to 100), this would clearly distort our analysis results. We can resolve this problem by standardizing the data prior to the analysis (Chap. 5).

Different standardization methods are available, such as *z*-standardization, which rescales each variable to a mean of 0 and a standard deviation of 1 (Chap. 5). In cluster analysis, however, *range standardization* (e.g., to a range of 0 to 1) typically works better (Milligan and Cooper 1988).

9.3.3.2 Binary and Nominal Variables

Whereas the distance measures presented thus far can be used for variables measured on a metric and, in general, on an ordinal scale, applying them to binary and nominal variables is problematic. When nominal variables are involved, you should rather select a similarity measure expressing the degree to which the variables' values share the same category. These **matching coefficients** can take different forms, but rely on the same allocation scheme as shown in Table 9.9. In this crosstab, cell *a* is the number of characteristics present in both objects, whereas cell *d* describes the number of characteristics absent in both objects. Cells *b* and *c* describe the number of characteristics present in one, but not the other, object (see Table 9.10 for an example).

The allocation scheme in Table 9.9 applies to binary variables (i.e., nominal variables with two categories). For nominal variables with more than two categories, you need to convert the categorical variable into a set of binary variables in order to use matching coefficients. For example, a variable with three categories needs to be transformed into three binary variables, one for each category (see the following example).

Based on the allocation scheme in Table 9.9, we can compute different matching coefficients, such as the **simple matching (SM) coefficient** (called *Matching* in Stata):

Table 9.9 Allocation scheme for matching coefficients

		Second object	
		Presence of a characteristics (1)	Absence of a characteristic (0)
First object	Presence of a characteristic (1)	a	b
	Absence of a characteristic (0)	c	d

Table 9.10 Recoded measurement data

Object	*Gender* (binary)		*Customer* (binary)		*Country of residence* (binary)		
	Male	Female	Yes	No	GER	UK	USA
A	1	0	1	0	1	0	0
B	1	0	0	1	0	0	1
C	0	1	0	1	0	0	1

$$SM = \frac{a + d}{a + b + c + d}$$

This coefficient takes both the joint presence and the joint absence of a characteristic (as indicated by cells *a* and *d* in Table 9.9) into account. This feature makes the simple matching coefficient particularly useful for symmetric variables where the joint presence and absence of a characteristic carry an equal degree of information. For example, the binary variable *gender* has the possible states "male" and "female." Both are equally valuable and carry the same weight when the simple matching coefficient is computed. However, when the outcomes of a binary variable are not equally important (i.e., the variable is asymmetric), the simple matching coefficient proves problematic. An example of an asymmetric variable is the presence, or absence, of a relatively rare attribute, such as customer complaints. While you say that two customers who complained have something in common, you cannot say that customers who did not complain have something in common. The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent). The agreement of two 1s (i.e., a positive match) is more significant than the agreement of two 0s (i.e., a negative match). Similarly, the simple matching coefficient proves problematic when used on nominal variables with many categories. In this case, objects may appear very similar, because they have many negative matches rather than positive matches.

Given this issue, researchers have proposed several other matching coefficients, such as the **Jaccard coefficient (JC)** and the **Russell and Rao coefficient** (**RR**, called *Russell* in Stata), which (partially) omit the *d* cell from the calculation. Like the simple matching coefficient, these coefficients range from 0 to 1 with higher values indicating a greater degree of similarity.[4] They are defined as follows:

$$JC = \frac{a}{a + b + c}$$

$$RR = \frac{a}{a + b + c + d}$$

To provide an example that compares the three coefficients, consider the following three variables:

- *gender*: male, female
- *customer*: yes, no
- *country of residence*: GER, UK, USA

[4] There are many other matching coefficients, such as Yule's Q, Kulczynski, or Ochiai, which are also menu-accessible in Stata. However, since most applications of cluster analysis rely on metric or ordinal data, we will not discuss these. See Wedel and Kamakura (2000) for more information on alternative matching coefficients.

We first transform the measurement data into binary data by recoding the original three variables into seven binary variables (i.e., two for *gender* and *customer*; three for *country of residence*). Table 9.10 shows a binary data matrix for three objects A, B, and C. Object A is a male customer from Germany; object B is a male non-customer from the United States; object C is a female non-customer, also from the United States.

Using the allocation scheme from Table 9.9 to compare objects A and B yields the following results for the cells: $a = 1$, $b = 2$, $c=2$, and $d = 2$.

This means that the two objects have only one shared characteristic ($a = 1$), but two characteristics, which are absent from both objects ($d = 2$). Using this information, we can now compute the three coefficients described earlier:

$$SM(A,B) = \frac{1+2}{1+2+2+2} = 0.571,$$

$$JC(A,B) = \frac{1}{1+2+2} = 0.2, \text{ and}$$

$$RR(A,B) = \frac{1}{1+2+2+2} = 0.143$$

As can be seen, the simple matching coefficient suggests that objects A and B are reasonably similar. Conversely, the Jaccard coefficient, and particularly the Russel Rao coefficient, suggests that they are not.

Try computing the distances between the other object pairs. Your computation should yield the following: $SM(A,C) = 0.143$, $SM(B,C) = 0.714$, $JC(A,C) = 0$, $JC(B,C) = 0.5$, $RR(A,C) = 0$, and $RR(B,C) = 0.286$.

9.3.3.3 Mixed Variables

Most datasets contain variables that are measured on multiple scales. For example, a market research questionnaire may require the respondent's gender, income category, and age. We therefore have to consider variables measured on a nominal, ordinal, and metric scale. How can we simultaneously incorporate these variables into an analysis?

A common approach is to dichotomize all the variables and apply the matching coefficients discussed above. For metric variables, this involves specifying categories (e.g., low, medium, and high age) and converting these into sets of binary variables. In most cases, the specification of categories is somewhat arbitrary. Furthermore, this procedure leads to a severe loss in precision, as we disregard more detailed information on each object. For example, we lose precise information on each respondent's age when scaling this variable down into age categories.

Gower (1971) introduced a dissimilarity coefficient that works with a mix of binary and continuous variablesa. **Gower's dissimilarity coefficient** is a composite measure that combines several measures into one, depending on each variable's scale level. If binary variables are used, the coefficient takes the value 1 when two

Table 9.11 Recoded measurement data

Object	*Gender* (binary)		*Customer* (binary)		*Income category* (ordinal)	*Age* (metric)
	Male	Female	Yes	No		
A	1	0	1	0	2	21
B	1	0	0	1	3	37
C	0	1	0	1	1	29

objects do not share a certain characteristic (cells *b* and *c* in Table 9.9), and 0 else (cells *a* and *d* in Table 9.9). Thus, when all the variables are binary and symmetric, Gower's dissimilarity coefficient reduces to the simple matching coefficient when expressed as a distance measure instead of a similarity measure (i.e., 1 – *SM*). If binary and asymmetric variables are used, Gower's dissimilarity coefficient equals the Jaccard coefficient when expressed as a distance measure instead of a similarity measure (i.e., 1 – *JC*). If continuous variables are used, the coefficient is equal to the city-block distance divided by each variable's range. Ordinal variables are treated as if they were continuous, which is fine when the scale is equidistant (see Chap. 3). Gower's dissimilarity coefficient welds the measures used for binary and continuous variables into one value that is an overall measure of dissimilarity.

To illustrate Gower's dissimilarity coefficient, consider the following example with the two binary variables *gender* and *customer*, the ordinal variable *income category* (1 = "low", 2 = "medium", 3 = "high"), and the metric variable *age*. Table 9.11 shows the data for three objects A, B, and C.

To compute Gower's dissimilarity coefficient for objects A and B, we first consider the variable *gender*. Since both objects A and B are male, they share two characteristics (male = "yes", female = "no"), which entails a distance of 0 for both variable levels. With regard to the *customer* variable, the two objects have different characteristics, hence a distance of 1 for each variable level. The ordinal variable income category is treated as continuous, using the city-block distance (here: |2–3|) divided by the variable's range (here: 3 – 1). Finally, the distance with regard to the *age* variable is |21 – 37|/(37 – 21) = 1. Hence, the resulting Gower distance is:

$$d_{Gower}(A,B) = \frac{1}{6}(0+0+1+1+0.5+1) \approx 0.583$$

Computing the Gower distance between the other two object pairs yields $d_{Gower}(A,C) \approx 0.833$, and $d_{Gower}(B,C) \approx 0.583$.

9.3.4 Decide on the Number of Clusters

An important question we haven't yet addressed is how to decide on the number of clusters. A misspecified number of clusters results in under- or oversegmentation, which easily leads to inaccurate management decisions on, for example, customer targeting, product positioning, or determining the optimal marketing mix (Becker et al. 2015).

We can select the number of clusters pragmatically, choosing a grouping that "works" for our analysis, but sometimes we want to select the "best" solution that the data suggest. However, different clustering methods require different approaches to decide on the number of clusters. Hence, we discuss hierarchical and portioning methods separately.

9.3.4.1 Hierarchical Methods: Deciding on the Number of Clusters

To guide this decision, we can draw on the distances at which the objects were combined. More precisely, we can seek a solution in which an additional combination of clusters or objects would occur at a greatly increased distance. This raises the issue of what a great distance is.

We can seek an answer by plotting the distance level at which the mergers of objects and clusters occur by using a **dendrogram**. Figure 9.15 shows the dendrogram for our example as produced by Stata. We read the dendrogram from the bottom to the top. The horizontal lines indicate the distances at which the objects were merged. For example, according to our calculations above, objects B and C were merged at a distance of 21.260. In the dendrogram, the horizontal line linking the two vertical lines that go from B and C indicates this merger. To decide on the number of clusters, we cut the dendrogram horizontally in the area where no merger has occurred for a long distance. In our example, this is done when moving from a four-cluster solution, which occurs at a distance of 28.160 (Table 9.4), to a three-cluster solution, which occurs at a distance of 36.249 (Table 9.5). This result suggests a four-cluster solution [A,D], [B,C,E], [F], and [G], but this conclusion is not clear-cut. In fact, the dendrogram often does not provide a clear indication, because it is generally difficult to identify where the cut should be made. This is particularly true of large sample sizes when the dendrogram becomes unwieldy.

Research has produced several other criteria for determining the number of clusters in a dataset (referred to as *stopping rules* in Stata).[5] One of the most prominent criteria is Calinski and Harabasz's (1974) **variance ratio criterion** (**VRC**; also called *Calinski-Harabasz pseudo-F* in Stata). For a solution with n objects and k clusters, the VRC is defined as:

$$VRC_k = (SS_B/(k-1))/(SS_W/(n-k)),$$

where SS_B is the sum of the squares between the clusters and SS_W is the sum of the squares within the clusters. The criterion should seem familiar, as it is similar to the F-value of a one-way ANOVA (see Chap. 6). To determine the appropriate number of clusters, you should choose the number that maximizes the VRC. However, as the VRC usually decreases with a greater number of clusters, you should compute

[5] For details on the implementation of these stopping rules in Stata, see Halpin (2016).

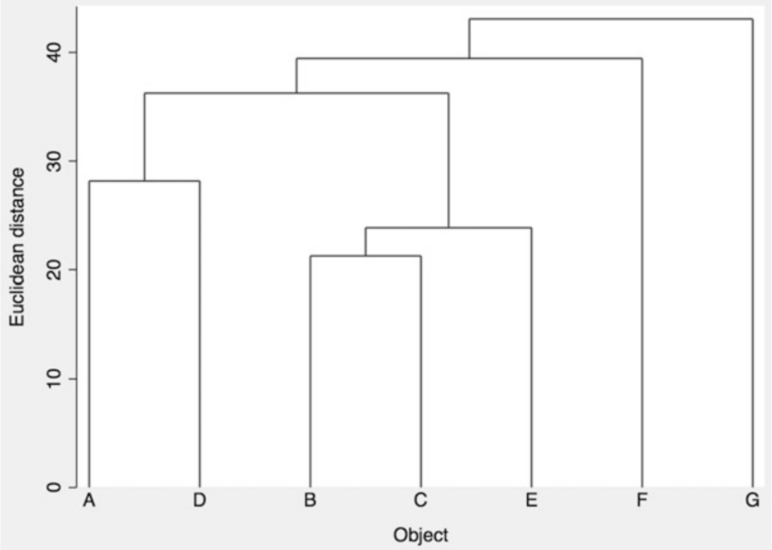


Fig. 9.15 Dendrogram

the difference in the VRC values ω_k of each cluster solution, using the following formula:[6]

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}).$$

The number of clusters k that minimizes the value in ω_k indicates the best cluster solution. Prior research has shown that the VRC reliably identifies the correct number of clusters across a broad range of constellations (Miligan and Cooper 1985). However, owing to the term VRC_{k-1}, which is not defined for a one-cluster solution, the minimum number of clusters that can be selected is three, which is a disadvantage when using the ω_k statistic.

Another criterion, which works well for determining the number of clusters (see Miligan and Cooper 1985) is the **Duda-Hart index** (Duda and Hart 1973). This index essentially performs the same calculation as the VRC, but compares the SS_W values in a pair of clusters to be split both before and after this split. More precisely, the Duda-Hart index is the SS_W in the two clusters (*Je(2)*) divided by the SS_W in one cluster (*Je(1)*); that is:

$$Duda - Hart = \frac{Je(2)}{Je(1)}$$

[6] In the ⭳ Web Appendix (→Downloads), we offer a Stata.ado file to calculate the ω_k called chomega.ado. We also offer an Excel sheet (*VRC.xlsx*) to calculate the ω_k manually.

To determine the number of clusters, you should choose the solution, which *maximizes* the Je(2)/Je(1) index value.

Duda et al. (2001) have also proposed a modified version of the index, which is called the *pseudo T-squared*. This index takes the number of observations in both groups into account. Contrary to the Duda-Hart index, you should choose the number of clusters that *minimizes* the pseudo T-squared.

Two aspects are important when using the Duda-Hart indices:

- The indices are not appropriate in combination with single linkage clustering, as chaining effects may occur. In this case, both indices will produce ambiguous results, as evidenced in highly similar values for different cluster solutions (Everitt and Rabe-Hesketh 2006).
- The indices are considered "local" in that they do not consider the entire data structure in their computation, but only the SS_W in the group being split. With regard to our example above, the Je(2)/Je(1) index for a two-cluster solution would only consider the variation in objects A to F, but not G. This characteristic makes the Duda-Hart indices somewhat inferior to the VRC, which takes the entire variation into account (i.e., the criterion is "global").

In practice, you should combine the VRC and the Duda-Hart indices by selecting the number of clusters that yields a large VRC, a large Je(2)/Je (1) index, and a small pseudo T-squared value. These values do not necessarily have to be the maximum or minimum values. Note that the VRC and Duda-Hart indices become less informative as the number of objects in the clusters becomes smaller.

Overall, the above criteria can often only provide rough guidance regarding the number of clusters that should be selected; consequently, you should instead take practical considerations into account. Occasionally, you might have a priori knowledge, or a theory on which you can base your choice. However, first and foremost, you should ensure that your results are interpretable and meaningful. Not only must the number of clusters be small enough to ensure manageability, but each segment should also be large enough to warrant strategic attention.

9.3.4.2 Partitioning Methods: Deciding on the Number of Clusters

When running partitioning methods, such as *k*-means, you have to pre-specify the number of clusters to retain from the data. There are varying ways of guiding this decision:

- Compute the VRC (see discussion in the context of hierarchical clustering) for an alternating number of clusters and select the solution that maximizes the VRC or minimizes ω_k. For example, compute the VRC for a three- to five-cluster

solution and select the number of clusters that minimizes ω_k. Note that the Duda-Hart indices are not applicable as they require a hierarchy of objects and mergers, which partitioning methods do not produce.
- Run a hierarchical procedure to determine the number of clusters by using the dendrogram and run *k*-means afterwards.[7] This approach also enables you to find starting values for the initial cluster centers to handle a second problem, which relates to the procedure's sensitivity to the initial classification (we will follow this approach in the example application).
- Rely on prior information, such as earlier research findings.

9.3.5 Validate and Interpret the Clustering Solution

Before interpreting the cluster solution, we need to assess the stability of the results. Stability means that the cluster membership of individuals does not change, or only changes a little when different clustering methods are used to cluster the objects. Thus, when different methods produce similar results, we claim stability.

The aim of any cluster analysis is to differentiate well between the objects. The identified clusters should therefore differ substantially from each other and the members of different clusters should respond differently to different marketing-mix elements and programs.

Lastly, we need to profile the cluster solution by using observable variables. **Profiling** ensures that we can easily assign new objects to clusters based on observable traits. For example, we could identify clusters based on loyalty to a product, but in order to use these different clusters, their membership should be identifiable according to tangible variables, such as income, location, or family size, in order to be actionable.

The key to successful segmentation is to critically revisit the results of different cluster analysis set-ups (e.g., by using different algorithms on the same data) in terms of managerial relevance. The following criteria help identify a clustering solution (Kotler and Keller 2015; Tonks 2009).

- *Substantial*: The clusters are large and sufficiently profitable to serve.
- *Reliable*: Only clusters that are stable over time can provide the necessary basis for a successful marketing strategy. If clusters change their composition quickly, or their members' behavior, targeting strategies are not likely to succeed. Therefore, a certain degree of stability is necessary to ensure that marketing strategies can be implemented and produce adequate results. Reliability can be evaluated by critically revisiting and replicating the clustering results at a later date.
- *Accessible*: The clusters can be effectively reached and served.

[7] See Punj and Stewart (1983) for additional information on this sequential approach.

- *Actionable*: Effective programs can be formulated to attract and serve the clusters.
- *Parsimonious*: To be managerially meaningful, only a small set of substantial clusters should be identified.
- *Familiar*: To ensure management acceptance, the cluster composition should be easy to relate to.
- *Relevant*: Clusters should be relevant in respect of the company's competencies and objectives.

9.3.5.1 Stability

Stability is evaluated by using different clustering procedures on the same data and considering the differences that occur. For example, you may first run a hierarchical clustering procedure, followed by *k*-means clustering to check whether the cluster affiliations of the objects change. Alternatively, running a hierarchical clustering procedure, you can use different distance measures and evaluate their effect on the stability of the results. However, note that it is common for results to change even when your solution is adequate. As a rule of thumb, if more than 20% of the cluster affiliations change from one technique to the other, you should reconsider the analysis and use, for example, a different set of clustering variables, or reconsider the number of clusters. Note, however, that this percentage is likely to increase with the number of clusters used.

When the data matrix exhibits identical values (referred to as **ties**), the ordering of the objects in the dataset can influence the results of the hierarchical clustering procedure. For example, the distance matrix based on the city-block distance in Table 9.8 shows the distance of 56 for object pairs (D,E), (E,F), and (F,G). Ties can prove problematic when they occur for the minimum distance in a distance matrix, as the decision about which objects to merge then becomes ambiguous (i.e., should we merge objects D and E, E and F, or F and G if 56 was the smallest distance in the matrix?). To handle this problem, van der Kloot et al. (2005) recommend re-running the analysis with a different input order of the data. The downside of this approach is that the labels of a cluster may change from one analysis to the next. This issue is referred to as *label switching*. For example, in the first analysis, cluster 1 may correspond to cluster 2 in the second analysis. Ties are, however, more the exception than the rule in practical applications—especially when using (squared) Euclidean distances—and generally don't have a pronounced impact on the results. However, if changing the order of the objects also drastically changes the cluster compositions (e.g., in terms of cluster sizes), you should reconsider the set-up of the analysis and, for example, re-run it with different clustering variables.

9.3.5.2 Differentiation of the Data

To examine whether the final partition differentiates the data well, we need to examine the cluster centroids. This step is highly important, as the analysis sheds light on whether the clusters are truly distinct. Only if objects across two (or more) clusters exhibit significantly different means in the clustering variables (or any other relevant variable) can they be distinguished from each other. This can be easily ascertained by comparing the means of the clustering variables across the clusters with independent *t*-tests or ANOVA (see Chap. 6).

Furthermore, we need to assess the solution's criterion validity. We do this by focusing on the criterion variables that have a theoretical relationship with the clustering variables, but were not included in the analysis. In market research, criterion variables are usually managerial outcomes, such as the sales per person, or willingness-to-pay. If these criterion variables differ significantly, we can conclude that the clusters are distinct groups with criterion validity.

9.3.5.3 Profiling

As indicated at the beginning of the chapter, cluster analysis usually builds on unobservable clustering variables. This creates an important problem when working with the final solution: How can we decide to which cluster a new object should be assigned if its unobservable characteristics, such as personality traits, personal values, or lifestyles, are unknown? We could survey these attributes and make a decision based on the clustering variables. However, this is costly and researchers therefore usually try to identify observable variables (e.g., demographics) that best mirror the partition of the objects. More precisely, these observable variables should partition the data into similar groups as the clustering variables do. Using these observable variables, it is then easy to assign a new object (whose cluster membership is unknown) to a certain cluster. For example, assume that we used a set of questions to assess the respondents' values and learned that a certain cluster comprises respondents who appreciate self-fulfillment, enjoyment of life, and a sense of accomplishment, whereas this is not the case in another cluster. If we were able to identify explanatory variables, such as gender or age, which distinguish these clusters adequately, then we could assign a new person to a specific cluster on the basis of these observable variables whose value traits may still be unknown.

9.3.5.4 Interpret the Clustering Solution

The interpretation of the solution requires characterizing each cluster by using the criterion or other variables (in most cases, demographics). This characterization should focus on criterion variables that convey why the cluster solution is relevant. For example, you could highlight that customers in one cluster have a lower willingness to pay and are satisfied with lower service levels, whereas customers in another cluster are willing to pay more for a superior service. By using this information, we can also try to find a meaningful name or label for each cluster; that is, one that adequately reflects the objects in the cluster. This is usually a challenging task, especially when unobservable variables are involved.

> While companies develop their own market segments, they frequently use standardized segments, based on established buying trends, habits, and customers' needs to position their products in different markets. The

(continued)

PRIZM lifestyle by Nielsen is one of the most popular segmentation databases. It combines demographic, consumer behavior, and geographic data to help marketers identify, understand, and reach their customers and prospective customers. PRIZM defines every US household in terms of more than 60 distinct segments to help marketers discern these consumers' likes, dislikes, lifestyles, and purchase behaviors.

An example is the segment labeled "Connected Bohemians," which Nielsen characterizes as a "collection of mobile urbanites, Connected Bohemians represent the nation's most liberal lifestyles. Its residents are a progressive mix of tech savvy, young singles, couples, and families ranging from students to professionals. In their funky row houses and apartments, Bohemian Mixers are the early adopters who are quick to check out the latest movie, nightclub, laptop, and microbrew." Members of this segment are between 25 and 44 years old, have a midscale income, own a hybrid vehicle, eat at Starbucks, and go skiing/snowboarding. (http://www.MyBestSegments.com).

Table 9.12 summarizes the steps involved in a hierarchical and *k*-means clustering when using Stata. The syntax code shown in the cells comes from the case study, which we introduce in the following section.

Table 9.12 Steps involved in carrying out a cluster analysis in Stata

Theory	Action
Research problem	
Identification of homogenous groups of objects in a population	
Select clustering variables to form segments	Select relevant variables that potentially exhibit high degrees of criterion validity with regard to a specific managerial objective.
Requirements	
Sufficient sample size	Make sure that the relationship between the objects and the clustering variables is reasonable. Ten times the number of clustering variables is the bare minimum, but 30 to 70 times is recommended. Ensure that the sample size is large enough to guarantee substantial segments.
Low levels of collinearity among the variables	► Statistics ► Summaries, tables and tests ► Summary and descriptive statistics ► Pairwise correlations
	`pwcorr e1 e5 e9 e21 e22`
	In case of highly correlated variables (correlation coefficients > 0.90), delete one variable of the offending pair.
Specification	
Choose the clustering procedure	If there is a limited number of objects in your dataset, rather use hierarchical clustering:
	► Statistics ► Multivariate analysis ► Cluster analysis ► Cluster Data ► Choose a linkage algorithm
	`cluster wardslinkage e1 e5 e9 e21 e22, measure (L2squared) name(wards_linkage)`

(continued)

Table 9.12 (continued)

Theory	Action
	If there are many observations (> 500) in your dataset, rather use *k*-means clustering: ▶ Statistics Multivariate analysis ▶ Cluster analysis ▶ Cluster Data ▶ kmeans
	`cluster kmeans e1 e5 e9 e21 e22, k(2) measure (L2squared) start(krandom) name(kmeans)`
Select a measure of (dis) similarity	*Hierarchical methods*:
	Select from the (dis)similarity measure menu, depending on the clustering variables' scale level.
	Depending on the scale level, select the measure; convert variables with multiple categories into a set of binary variables and use matching coefficients; Choose Gower's dissimilarity coefficient for mixed variables.
	When the variables are measured on different units, standardize the variables to a range from 0 to 1 prior to the analysis, using the following commands:
	`summarize e1`
	`return list`
	`gen e1_rsdt =.`
	`replace e1_rsdt = (e1- r(min)) / (r(max)-r (min))`
	Partitioning methods:
	Use the squared Euclidean distance from the (dis) similarity menu.
Deciding on the number of clusters	*Hierarchical clustering:*
	Examine the dendrogram:
	▶ Statistics ▶ Multivariate analysis ▶ Cluster analysis ▶ Postclustering ▶ Dendrogram
	`cluster dendrogram wards_linkage, cutnumber (10) showcount`
	Examine the VRC and Duda-Hart indices:
	▶ Statistics Multivariate analysis ▶ Cluster analysis ▶ Postclustering ▶ Cluster analysis stopping rules.
	For VRC: `cluster stop wards_linkage, rule (calinski) groups(2/11)`
	For Duda-Hart: `cluster stop wards_linkage, rule (duda) groups(1/10)`
	Include practical considerations in your decision.
	Partitioning methods:
	Run a hierarchical cluster analysis and decide on the number of segments based on a dendrogram, the VRC, and the Duda-Hart indices; use the resulting partition as starting partition.
	▶ Statistics Multivariate analysis ▶ Cluster analysis ▶ Postclustering ▶ Cluster analysis stopping rules.
	`cluster kmeans e1 e5 e9 e21 e22, k(3) measure (L2squared) name(kmeans) start(group (cluster_wl))`
	Include practical considerations in your decision.

(continued)

Table 9.12 (continued)

Theory	Action
Validating and interpreting the cluster solution	
Stability	Re-run the analysis using different clustering procedures, linkage algorithms or distance measures. For example, generate a cluster membership variable and use this grouping as starting partition for *k*-means clustering.
	`cluster generate cluster_wl = groups(3), name (wards_linkage) ties(error)`
	`cluster kmeans e1 e5 e9 e21 e22, k(3) measure (L2squared) name (kmeans) start(group (cluster_wl))`
	Examine the overlap in the clustering solutions. If more than 20% of the cluster affiliations change from one technique to the other, you should reconsider the set-up.
	`tabulate cluster_wl kmeans`
	Change the order of objects in the dataset (hierarchical clustering only).
Differentiation of the data	Compare the cluster centroids across the different clusters for significant differences.
	`mean e1 e5 e9 e21 e22, over(cluster_wl)`
	If possible, assess the solution's criterion validity.
Profiling	Identify observable variables (e.g., demographics) that best mirror the partition of the objects based on the clustering variables.
	`tabulate cluster_wl flight_purpose, chi2 V`
Interpretating of the cluster solution	Identify names or labels for each cluster and characterize each cluster by means of observable variables.

9.4 Example

Let's go back to the Oddjob Airways case study and run a cluster analysis on the data. Our aim is to identify a manageable number of segments that differentiates the customer base well. To do so, we first select a set of clustering variables, taking the sample size and potential collinearity issues into account. Next, we apply hierarchical clustering based on the squared Euclidean distances, using the Ward's linkage algorithm. This analysis will help us determine a suitable number of segments and a starting partition, which we will then use as the input for *k*-means clustering.

9.4.1 Select the Clustering Variables

The Oddjob Airways dataset (⭳ Web Appendix → Downloads) offers several variables for segmenting its customer base. Our analysis draws on the following set of variables, which we consider promising for identifying distinct segments based on customers' expectations regarding the airline's service quality (variable names in parentheses):

- . . . with Oddjob Airways you will arrive on time (*e1*),
- . . . Oddjob Airways provides you with a very pleasant travel experience (*e5*),
- . . . Oddjob Airways gives you a sense of safety (*e9*),
- . . . Oddjob Airways makes traveling uncomplicated (*e21*), and
- . . . Oddjob Airways provides you with interesting on-board entertainment, service, and information sources (*e22*).

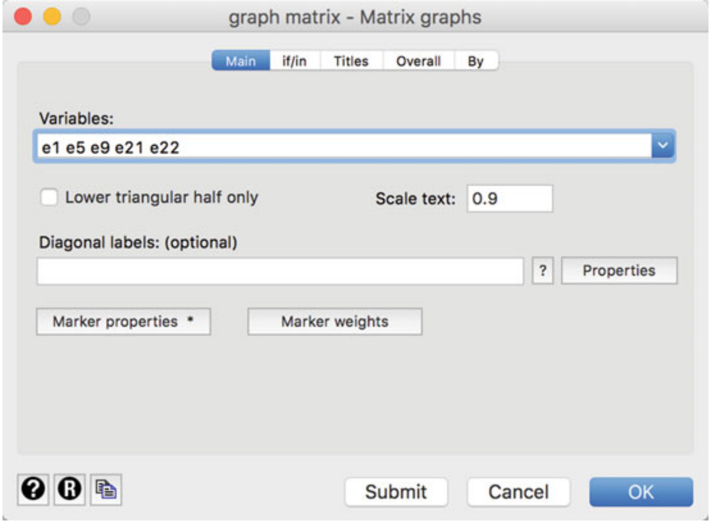
With five clustering variables, our analysis meets even the most conservative rule-of-thumb regarding minimum sample size requirements. Specifically, according to Dolnicar et al. (2016), the cluster analysis should draw on 100 times the number of clustering variables to optimize cluster recovery. As our sample size of 1,065 is clearly higher than $5 \cdot 100 = 500$, we can proceed with the analysis. Note, however, that the actual sample size used in the analysis may be substantially lower when using casewise deletion. This also applies to our analysis, which ultimately draws on 969 observations (i.e., after casewise deletion).

To begin with, it is good practice to examine a graphical display of the data. With multivariate data such as ours, the best way to visualize the data is by means of a scatterplot matrix (see Chaps. 5 and 7). To generate a scatterplot matrix, go to ► Graphics ► Scatterplot matrix and enter the variables *e1*, *e5*, *e9*, *e21*, and *e22* into the **Variables** box (Fig. 9.16). To ensure that the variable labels fit the diagonal boxes of the scatterplot, enter **0.9** next to **Scale text**. Because there are so many observations in the dataset, we choose a different marker symbol. To do so, click on **Marker properties** and select **Point** next to **Symbol**. Confirm by clicking on **Accept**, followed by **OK**. Stata will generate a scatterplot similar to the one shown in Fig. 9.17.

The resulting scatterplots do not suggest a clear pattern except that most observations are in the moderate to high range. But the scatterplots also assure us that all observations fall into the 0 to 100 range. Even though some observations with low values in (combinations of) expectation variables can be considered as extreme, we do not delete them, as they occur naturally in the dataset (see Chap. 5).

In a further check, we examine the variable correlations by clicking on ► Statistics ► Summaries, tables and tests ► Summary and descriptive statistics ► Pairwise correlations. Next, enter all variables into the **Variables** box (Fig. 9.18). Click on **OK** and Stata will display the results (Table 9.13).

The results show that collinearity is not at a critical level. The variables *e1* and *e21* show the highest correlation of 0.6132, which is clearly lower than the 0.90 threshold. We can therefore proceed with the analysis, using all five clustering variables.



graph matrix - Matrix graphs

Variables:
e1 e5 e9 e21 e22

□ Lower triangular half only

Scale text: 0.9

Diagonal labels: (optional)

Fig. 9.16 Scatterplot matrix dialog box

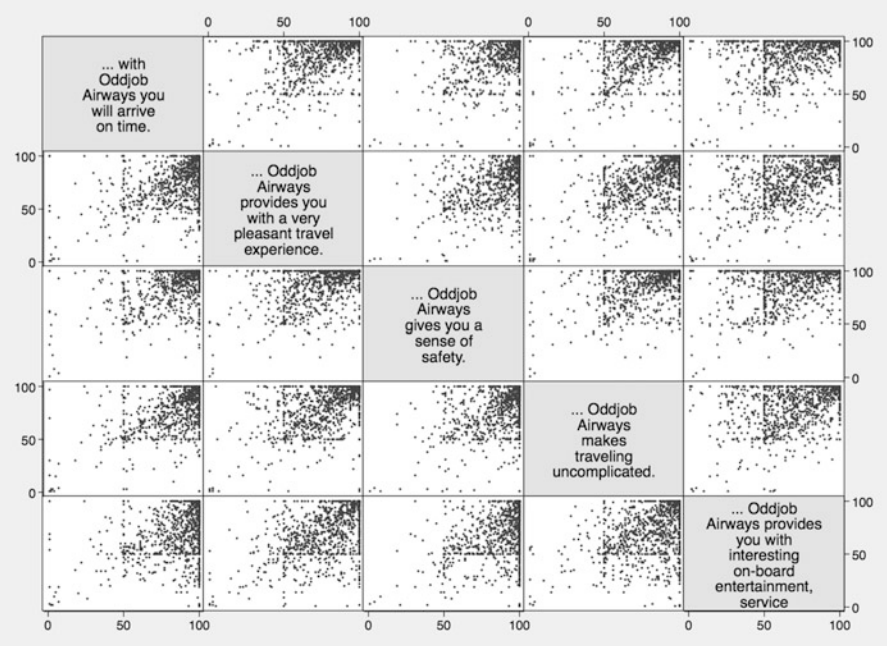
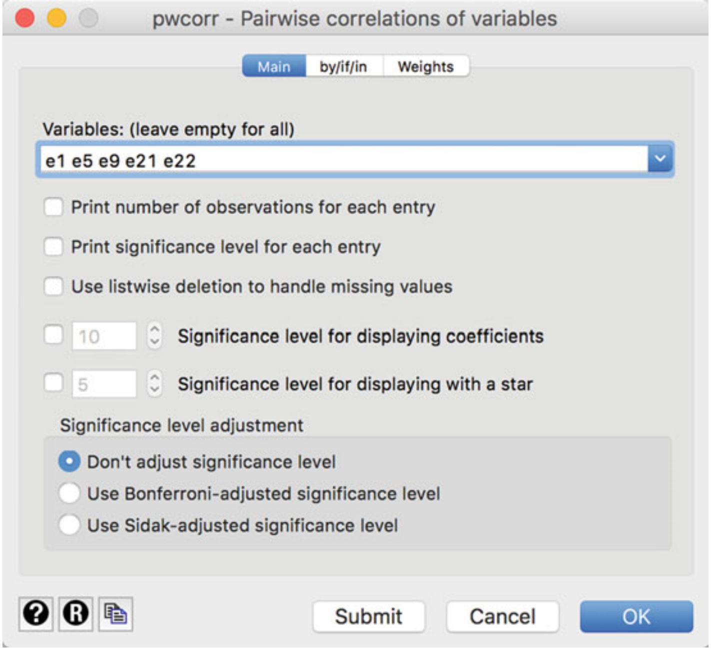


Fig. 9.17 Scatterplot matrix

Note that all the variables used in our analysis are metric and are measured on a scale from 0 to 100. However, if the variables were measured in different units with



Variables: (leave empty for all)
e1 e5 e9 e21 e22

□ Print number of observations for each entry

□ Print significance level for each entry

□ Use listwise deletion to handle missing values

□ 10 Significance level for displaying coefficients

□ 5 Significance level for displaying with a star

Significance level adjustment

● Don't adjust significance level

○ Use Bonferroni-adjusted significance level

○ Use Sidak-adjusted significance level

Fig. 9.18 Pairwise correlations dialog box

Table 9.13 Pairwise correlations

```
pwcorr e1 e5 e9 e21 e22

             |       e1       e5       e9      e21      e22
-------------+---------------------------------------------
          e1 |   1.0000
          e5 |   0.5151   1.0000
          e9 |   0.5330   0.5255   1.0000
         e21 |   0.6132   0.5742   0.5221   1.0000
         e22 |   0.3700   0.5303   0.4167   0.4246   1.0000
```

different variances, we would need to standardize them in order to avoid the variables with the highest variances dominating the analysis. In Box 9.3, we explain how to standardize the data in Stata.

Box 9.3 Standardization in Stata
Stata's menu-based dialog boxes only allow for *z*-standardization (see Chap. 5), which you can access via ► Data ► Create or change data ► Create new variable (extended). In cluster analysis, however, the clustering variables should be standardized to a scale of 0 to 1. There is no menu option

(continued)

Box 9.3 (continued)

or command to do this directly in Stata, but we can improvise by using the `summarize` command. When using this command, Stata saves the minimum and maximum values of a certain variable as scalars. Stata refers to these scalars as *r(max)* and *r(min)*, which we can use to calculate new versions of the variables, standardized to a scale from 0 to 1. To run this procedure for the variable *e1* type in the following:

```
summarize e1
  Variable |        Obs         Mean    Std. Dev.    Min     Max
-----------+------------------------------------------------------
        e1 |      1,038     86.08189     19.3953     1       100
```

We can let Stata display the results of the `summarize` command by typing `return list` in the command window.

```
scalars:
                  r(N)  =  1038
              r(sum_w)  =  1038
               r(mean)  =  86.08188824662813
                r(Var)  =  376.177472998l067
                 r(sd)  =  19.395295125316
                r(min)  =  1
                r(max)  =  100
                r(sum)  =  89353
```

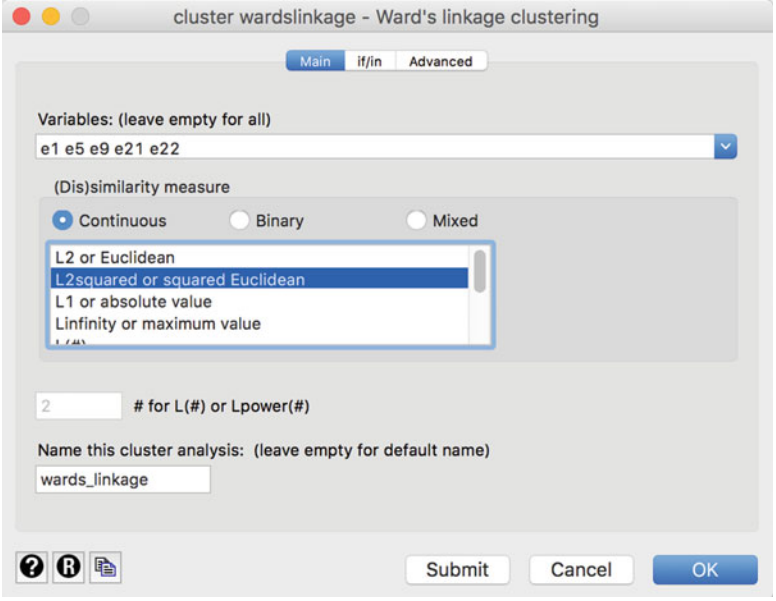
Next, we compute a new variable called *e1_rstd*, which uses the minimum and maximum values as input to compute a standardized version of *e1* (see Chap. 5 for the formula).

```
gen e1_rsdt =.
replace e1_rsdt = (e1- r(min)) / (r(max)-r(min))
```

Similar commands create standardized versions of the other clustering variables.

9.4.2 Select the Clustering Procedure and Measure of Similarity or Dissimilarity

To initiate hierarchical clustering, go to ► Statistics ► Multivariate analysis ► Cluster analysis ► Cluster data. The resulting menu offers a range of hierarchical and partitioning methods from which to choose. Because of its versatility and



cluster wardslinkage - Ward's linkage clustering

Variables: (leave empty for all)

e1 e5 e9 e21 e22

(Dis)similarity measure

● Continuous ○ Binary ○ Mixed

L2 or Euclidean
L2squared or squared Euclidean
L1 or absolute value
Linfinity or maximum value

2 # for L(#) or Lpower(#)

Name this cluster analysis: (leave empty for default name)

wards_linkage

Fig. 9.19 Hierarchical clustering with Ward's linkage dialog box

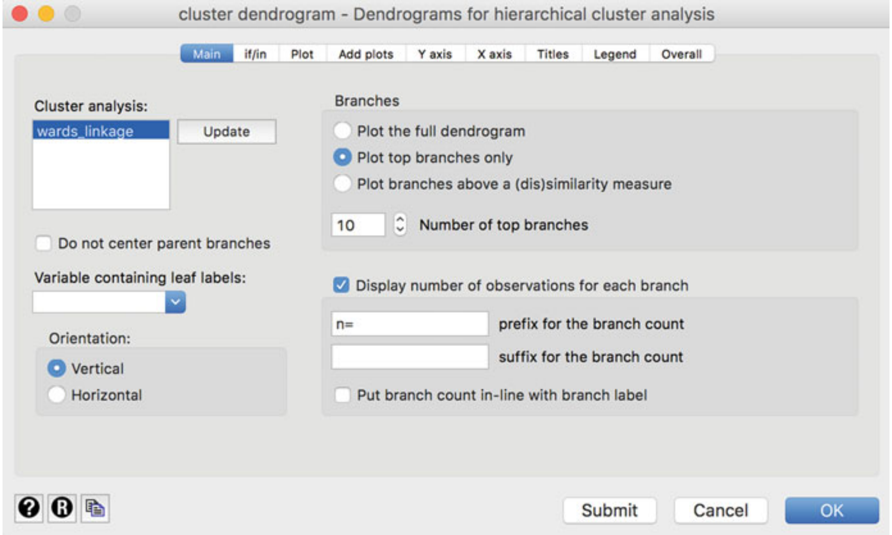
general performance, we choose **Ward's linkage**. Clicking on the corresponding menu option opens a dialog box similar to Fig. 9.19.

Enter the variables *e1*, *e5*, *e9*, *e21*, and *e22* into the **Variables** box and select the squared Euclidean distance (**L2squared or squared Euclidean**) as the (dis)similarity measure. Finally, specify a name, such as *wards_linkage*, in the **Name this cluster analysis** box. Next, click on **OK**.

Nothing seems to happen (aside from the following command, which gets issued: `cluster wardslinkage e1 e5 e9 e21 e22, measure (L2squared) name(wards_linkage)`), although you might notice that our dataset now contains three additional variables called *wards_linkage_id*, *wards_linkage_ord*, and *wards_linkage_hgt*. While these new variables are not directly of interest, Stata uses them as input to draw the dendrogram.

9.4.3 Decide on the Number of Clusters

To decide on the number of clusters, we start by examining the dendrogram. To display the dendrogram, go to ▶ Statistics ▶ Multivariate analysis ▶ Cluster analysis ▶ Postclustering ▶ Dendrogram. Given the great number of observations, we need to limit the display of the dendrogram (see Fig. 9.20). To do so, select **Plot top branches only** in the **Branches** menu. By specifying **10** next to **Number of branches**, we limit the view of the top 10 branches of the dendrogram, which Stata labels *G1* to *G10*. When selecting **Display number of observations for each branch**, Stata will display the number of observations in each of the ten groups.



cluster dendrogram - Dendrograms for hierarchical cluster analysis

Cluster analysis: wards_linkage

□ Do not center parent branches

Variable containing leaf labels:

Orientation:
● Vertical
○ Horizontal

Branches
○ Plot the full dendrogram
● Plot top branches only
○ Plot branches above a (dis)similarity measure

10 Number of top branches

☑ Display number of observations for each branch

n= prefix for the branch count

suffix for the branch count

□ Put branch count in-line with branch label

Fig. 9.20 Dendrogram dialog box

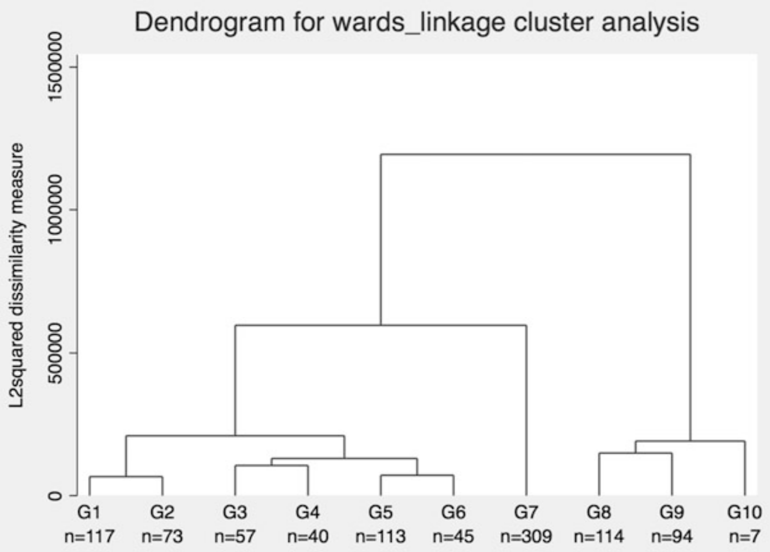
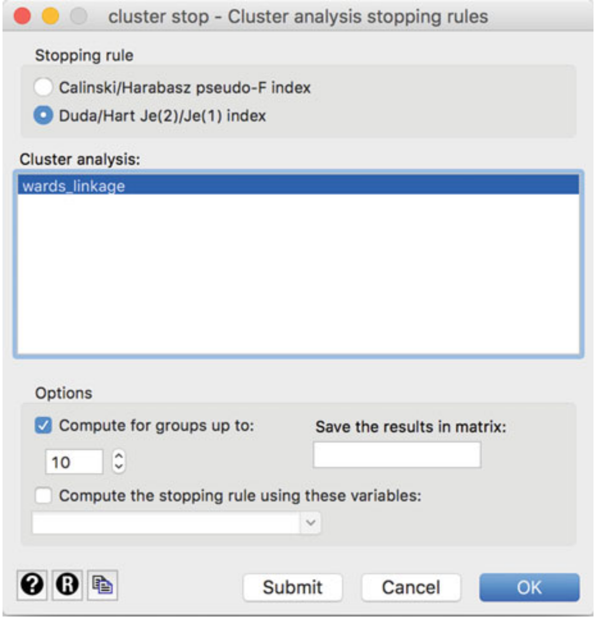


Fig. 9.21 Dendrogram

After clicking on **OK**, Stata will open a new window with the dendrogram (Fig. 9.21).

Reading the dendrogram from the bottom to the top, we see that clusters *G1* to *G6* are merged in quick succession. Clusters *G8* to *G10* are merged at about the


cluster stop - Cluster analysis stopping rules

Stopping rule

○ Calinski/Harabasz pseudo-F index

● Duda/Hart Je(2)/Je(1) index

Cluster analysis:

wards_linkage

Options

☑ Compute for groups up to:

10

Save the results in matrix:

□ Compute the stopping rule using these variables:

Fig. 9.22 Postclustering dialog box

same distance, while *G7* initially remains separate. These three clusters remain stable until, at a much higher distance, *G7* merges with the first cluster. This result clearly suggests a three-cluster solution, because reducing the cluster number to two requires merging the first cluster with *G7*, which is quite dissimilar to it. Increasing the number of clusters appears unreasonable, as many mergers take place at about the same distance.

The VRC and Duda-Hart indices allow us to further explore the number of clusters to extract. To request these measures in Stata, go to ► Statistics ► Multivariate analysis ► Cluster analysis ► Postclustering ► Cluster analysis stopping rules. In the dialog box that follows (Fig. 9.22), select **Duda/Hart Je(2)/J2(1) index** and tick the box next to **Compute for groups up to**. As we would like to consider a maximum number of ten clusters, enter **10** into the corresponding box and click on **OK**. Before interpreting the output, continue this procedure, but, this time, choose the **Calinski/Harabasz pseudo F-index** to request the VRC. Recall that we can also compute the VRC-based ω_k statistic. As this statistic requires the VRC$_{k+1}$ value as input, we need to enter **11** under **Compute for groups up to**. Next, click on **OK**. Tables 9.14 and 9.15 show the postclustering outputs.

Looking at Table 9.14, we see that the Je(2)/Je(1) index yields the highest value for three clusters (**0.8146**), followed by a six-cluster solution (**0.8111**). Conversely, the lowest pseudo T-squared value (**37.31**) occurs for ten clusters. Looking at the VRC values in Table 9.15, we see that the index decreases with a greater number of clusters.

To calculate the ω_k criterion, we can use a file that has been specially programmed for this, called chomega.ado ⭳ Web Appendix (→ Downloads). This

Table 9.14 Duda-Hart indices

```
cluster stop wards_linkage, rule(duda) groups(1/10)

+-----------------------------------------+
|             |         Duda/Hart         |
|  Number of  |             |  pseudo     |
|  clusters   | Je(2)/Je(1) |  T-squared  |
|-------------+-------------+-------------|
|      1      |   0.6955    |   423.29    |
|      2      |   0.6783    |   356.69    |
|      3      |   0.8146    |   100.83    |
|      4      |   0.7808    |    59.79    |
|      5      |   0.7785    |    58.59    |
|      6      |   0.8111    |    58.94    |
|      7      |   0.6652    |    47.82    |
|      8      |   0.7080    |    64.34    |
|      9      |   0.7127    |    75.79    |
|     10      |   0.7501    |    37.31    |
+-----------------------------------------+
```

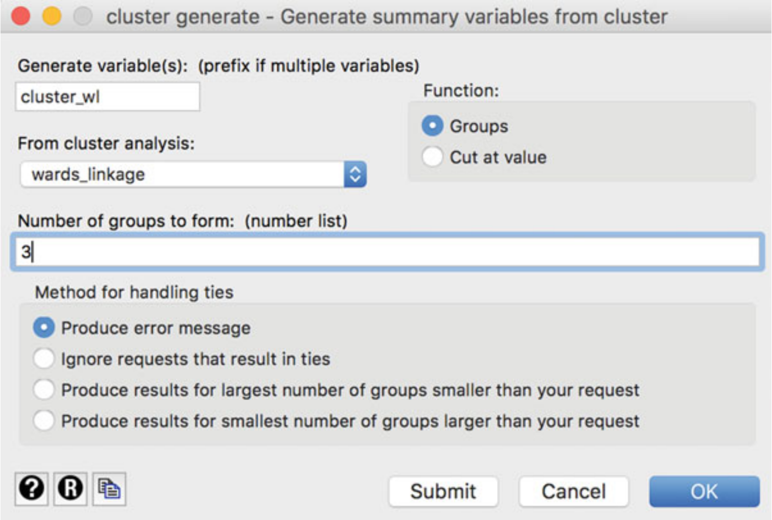
Table 9.15 VRC

```
cluster stop wards_linkage, rule(calinski) groups(2/10)

+---------------------------+
|             |  Calinski/  |
|  Number of  |  Harabasz   |
|  clusters   |  pseudo-F   |
|-------------+-------------|
|      2      |   423.29    |
|      3      |   406.02    |
|      4      |   335.05    |
|      5      |   305.39    |
|      6      |   285.26    |
|      7      |   273.24    |
|      8      |   263.12    |
|      9      |   249.61    |
|     10      |   239.73    |
|     11      |   233.17    |
+---------------------------+
```

Table 9.16 ω_k statistic

```
chomega
omega_3  is -53.691
omega_4  is  41.300
omega_5  is   9.534
omega_6  is   8.110
omega_7  is   1.899
omega_8  is  -3.394
omega_9  is   3.636
Minimum value of omega: -53.691 at 3 clusters
```

file should first be run before we can use it, just like the add-on modules discussed in Chap. 5, Section 5.8.2. To do this, download the chomega.ado file and drag it into the Stata command box, and add `do "` before and `"` after the text that appears in the



cluster generate - Generate summary variables from cluster

Generate variable(s): (prefix if multiple variables)
cluster_wl

From cluster analysis:
wards_linkage

Function:
● Groups
○ Cut at value

Number of groups to form: (number list)
3

Method for handling ties
● Produce error message
○ Ignore requests that result in ties
○ Produce results for largest number of groups smaller than your request
○ Produce results for smallest number of groups larger than your request

Fig. 9.23 Summary variables dialog box

Stata **Command** window (see Chap. 5). Then click on *enter*. Then you should type `chomega`. Note that this only works if you have first performed a cluster analysis.

The output is included in Table 9.16. We find that the smallest ω_k value of −**53.691** occurs for a three-cluster solution. The smallest value is shown at the top and bottom of Table 9.16. Note again that since ω_k requires VRC$_{k\text{-}1}$ as input, the statistic is only defined for three or more clusters. Taken jointly, our analyses of the dendrogram, the Duda-Hart indices, and the VRC clearly suggest a three-cluster solution.

9.4.4 Validate and Interpret the Clustering Solution

In the next step, we create a cluster membership variable, which indicates the cluster to which each observation belongs. To do so, go to ► Statistics ► Multivariate analysis ► Cluster analysis ► Postclustering ► Generate summary variables from cluster. In the dialog box that opens (Fig. 9.23), enter a name, such as *cluster_wl*, for the variable to be created in the **Generate variable(s)** box. In the dropdown list **From cluster analysis**, we can choose on which previously run cluster analysis the cluster membership variable should be based. As this is our first analysis, we can only select *wards_linkage*. Finally, specify the number of clusters to extract (**3**) under **Number of groups to form** and proceed by clicking **OK**.

Stata generates a new variable *cluster_wl*, which indicates the group to which each observation belongs. We can now use this variable to describe and profile the clusters. In a first step, we would like to tabulate the number of observations in each

Table 9.17 Cluster sizes

```
tabulate cluster_wl, missing

 cluster_wl |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        445       41.78       41.78
          2 |        309       29.01       70.80
          3 |        215       20.19       90.99
          . |         96        9.01      100.00
------------+-----------------------------------
      Total |      1,065      100.00
```

Table 9.18 Comparison of means

```
tabstat e1 e5 e9 e21 e22, statistics( mean ) by(cluster_wl)

Summary statistics: mean
  by categories of: cluster_wl

cluster_wl |        e1        e5        e9       e21       e22
-----------+--------------------------------------------------
         1 |  92.39326  75.50562  89.74607   81.7191  62.33933
         2 |   97.1068  95.54693  97.50809  96.63754  92.84466
         3 |   59.4186  58.28372  71.62791  56.72558  58.03256
-----------+--------------------------------------------------
     Total |  86.57998  78.07534  88.20124  80.93086  71.11146
--------------------------------------------------------------
```

cluster by going to ▶ Statistics ▶ Summary, tables, and tests ▶ Frequency tables ▶ One-way table. Simply select *cluster_wl* in the drop-down menu under **Categorical variable**, tick the box next to **Treat missing values like other values** and click on **OK**. The output in Table 9.17 shows that the cluster analysis assigned **969** observations to the three segments; **96** observations are not assigned to any segment due to missing values. The first cluster comprises **445** observations, the second cluster **309** observations, and the third cluster **215** observations.

Next, we would like to compute the centroids of our clustering variables. To do so, go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Other tables ▶ Compact table of summary statistics and enter *e1 e5 e9 e21 e22* into the **Variables** box. Next, click on **Group statistics by variable** and select *cluster_wl* from the list. Under **Statistics to display**, tick the first box and select **Mean**, followed by **OK**. Table 9.18 shows the resulting output.

Comparing the variable means across the three clusters, we find that respondents in the first cluster strongly emphasize punctuality (e_1), while comfort (e_5) and, particularly, entertainment aspects (*e22*) are less important. Respondents in the second cluster have extremely high expectations regarding all five performance features, as evidenced in average values well above 90. Finally, respondents in the third cluster do not express high expectations in general, except in terms of security (e_9). Based on these results, we could label the first cluster “on-time is enough,” the

Table 9.19 Crosstab

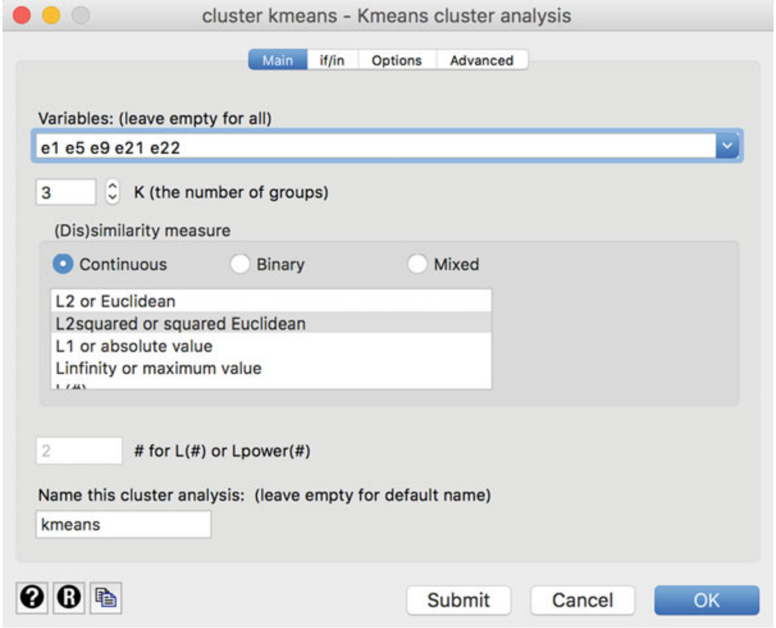
```
tabulate cluster_wl flight_purpose, chi2 V

           |  Do you normaly fly
           |    for business or
           |   leisure purposes?
cluster_wl |  Business   Leisure |     Total
-----------+----------------------+----------
         1 |       239       206 |       445
         2 |       130       179 |       309
         3 |       114       101 |       215
-----------+----------------------+----------
     Total |       483       486 |       969

          Pearson chi2(2) =  10.9943   Pr = 0.004
               Cramér's V =   0.1065
```

second cluster "the demanding traveler," and the third cluster "no thrills." We could further check whether these differences in means are significant by using a one-way ANOVA as described in Chap. 6.

In a further step, we can try to profile the clusters using sociodemographic variables. Specifically, we use crosstabs (see Chap. 5) to contrast our clustering with the variable *flight_purpose*, which indicates whether the respondents primarily fly for business purposes (*flight_purpose=1*) or private purposes (*flight_purpose=2*). To do so, click on ► Statistics ► Summaries, tables, and tests ► Frequency tables ► Two-way table with measures of association. In the dialog box that opens, enter *cluster_wl* into the **Row variable** box and *flight_purpose* into the **Column variable** box. Select **Pearson's chi-squared** and **Cramer's V** under **Test statistics** and click on **OK**. The results in Table 9.19 show that the majority of respondents in the first and third cluster are business travelers, whereas the second cluster primarily comprises private travelers. The χ^2-test statistic (**Pr** = **0.004**) indicates a significant relationship between these two variables. However, the strength of the variables' association is rather small, as indicated by the Cramer's V of **0.1065**.

The Oddjob Airways dataset offers various other variables such as *age*, *gender*, or *status*, which could be used to further profile the cluster solution. However, instead of testing these variables' efficacy step-by-step, we proceed and assess the solution's stability by running an alternative clustering procedure on the data. Specifically, we apply the *k*-means method, using the grouping from the Ward's linkage analysis as input for the starting partition. To do so, go to:

► Statistics ► Multivariate statistics ► Cluster analysis ► Cluster data ► Kmeans. In the dialog box that opens, enter *e1*, *e5*, *e9*, *e21*, and *e22* into the **Variables** box, choose **3** clusters, and select **L2squared or squared Euclidean** under **(Dis)similarity measure** (Fig. 9.24). Under **Name this cluster analysis**, make sure that you specify an intuitive name, such as *kmeans*. When clicking on the **Options** tab, we can choose between different options of how *k*-means should derive a starting partition for the analysis. Since we want to use the clustering from



Variables: (leave empty for all)
e1 e5 e9 e21 e22

3 K (the number of groups)

(Dis)similarity measure

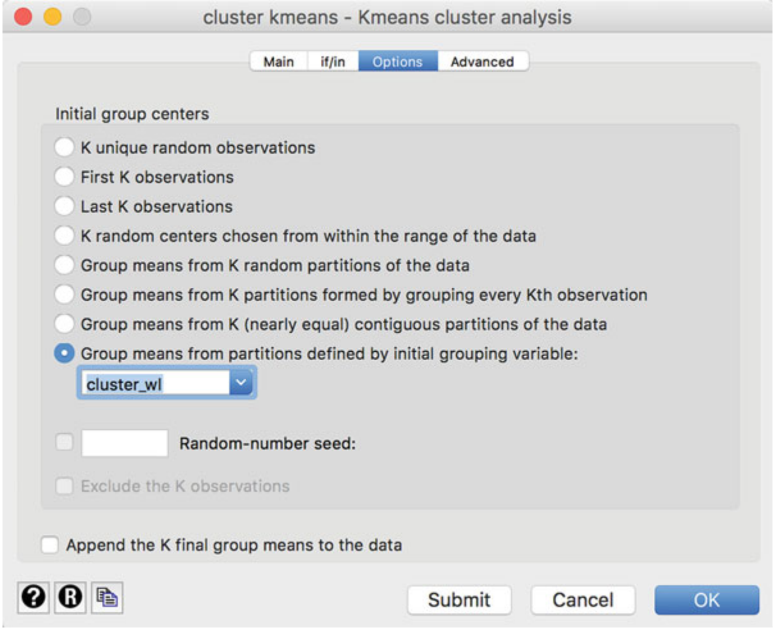
● Continuous ○ Binary ○ Mixed

L2 or Euclidean
L2squared or squared Euclidean
L1 or absolute value
Linfinity or maximum value

2 # for L(#) or Lpower(#)

Name this cluster analysis: (leave empty for default name)
kmeans

Fig. 9.24 *k*-means dialog box



Initial group centers

○ K unique random observations
○ First K observations
○ Last K observations
○ K random centers chosen from within the range of the data
○ Group means from K random partitions of the data
○ Group means from K partitions formed by grouping every Kth observation
○ Group means from K (nearly equal) contiguous partitions of the data
● Group means from partitions defined by initial grouping variable:
cluster_wl

□ Random-number seed:

□ Exclude the K observations

□ Append the K final group means to the data

Fig. 9.25 Options in the *k*-means dialog box

Table 9.20 Comparison of clustering results

```
tabulate cluster_wl kmeans


           |              kmeans
cluster_wl |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |       320        107         18 |       445
         2 |         2        307          0 |       309
         3 |        36         10        169 |       215
-----------+---------------------------------+----------
     Total |       358        424        187 |       969
```

our previous analysis by using Ward's linkage, we need to choose the last option and select the *cluster_wl* variable in the corresponding drop-down menu (Fig. 9.25). Now click on **OK**.

Stata only issues a command (`cluster kmeans e1 e5 e9 e21 e22, k(3) measure(L2squared) name(kmeans) start(group(cluster_wl))`) but also adds a new variable *kmeans* to the dataset, which indicates each observation's cluster affiliation, analogous to the *cluster_wl* variable for Ward's linkage. To explore the overlap in the two cluster solutions, we can contrast the results using crosstabs. To do so, go to ► Statistics ► Summary, tables, and tests ► Frequency tables ► Two-way table with measures of association and select *cluster_wl* under **Row variable** and *kmeans* under **Column variable**. After clicking on **OK**, Stata will produce an output similar to Table 9.20.

The results show that there is a strong degree of overlap between the two cluster analyses. For example, **307** observations that fall into the second cluster in the Ward's linkage analysis also fall into this cluster in the *k*-means clustering. Only two observations from this cluster appear in the first *k*-means cluster. The divergence in the clustering solutions is somewhat higher in the third and, especially, in the first cluster, but still low in absolute terms. Overall, the two analyses have an overlap of (320 + 307 + 169)/969 = 82.15%, which is very satisfactory as less than 20% of all observations appear in a different cluster when using *k*-means.

This analysis concludes our cluster analysis. However, we could further explore the solution's stability by running other linkage algorithms, such as centroid or complete linkage, on the data. Similarly, we could use different (dis)similarity measures and assess their impact on the results. So go ahead and explore these options yourself!

9.5 Oh, James! (Case Study)

The James Bond movie series is one of the success stories of filmmaking. The movies are the longest continually running and the third-highest-grossing film series to date, which started in 1962 with Dr. No, starring Sean Connery as James Bond. As of 2016, there have been 24 movies with six actors having played James

Bond. Interested in the factors that contributed to this running success, you decide to investigate the different James Bond movies' characteristics. Specifically, you want to find out whether the movies can be grouped into clusters, which differ in their box-office revenues. To do so, you draw on Internet Movie Database (www.imdb.com) and collect data on all 24 movies based on the following variables (variable names in parentheses):

- Title. (*title*)
- Actor playing James Bond. (*actor*)
- Year of publication. (*year*)
- Budget in USD, adjusted for inflation. (*budget*)
- Box-office revenues in the USA, adjusted for inflation. (*gross_usa*)
- Box-office revenues worldwide, adjusted for inflation. (*gross_worldwide*)
- Runtime in minutes. (*runtime*)
- Native country of the villain actor. (*villain_country*)
- Native country of the bondgirl. (*bondgirl_country*)
- Haircolor of the bondgirl. (*bondgirl_hair*)

Use the dataset *jamesbond.dta* (⭳ Web Appendix → Downloads) to run a cluster analysis—despite potential objections regarding the sample size. Answer the following questions:

1. Which clustering variables would you choose in light of the study objective, their levels of measurement, and correlations?
2. Given the levels of measurement, which clustering method would you prefer? Carry out a cluster analysis using this procedure.
3. Interpret and profile the obtained clusters by examining cluster centroids. Compare the differences across clusters on the box-office revenue variables.
4. Use a different clustering method to test the stability of your results.

9.6 Review Questions

1. In your own words, explain the objective and basic concept of cluster analysis.
2. What are the differences between hierarchical and partitioning methods? When do we use hierarchical or partitioning methods?
3. Repeat the manual calculations of the hierarchical clustering procedure from the beginning of the chapter, but use complete linkage as the clustering method. Compare the results with those of the single linkage method.
4. Explain the different options to decide on the number of clusters to extract from the data? Should you rather on statistical measures or rather on practical reasoning?
5. Run the *k*-means analysis on the Oddjob Airways data again (*oddjob.dta*, ⭳ Web Appendix → Downloads). Assume a three-cluster solution and try the different

options for obtaining a starting partition that Stata offers. Compare the results with those obtained by the hierarchical clustering.

6. Which clustering variables could be used to segment:
   – The market for smartphones?
   – The market for chocolate?
   – The market for car insurances?

9.7 Further Readings

Bottomley, P., & Nairn, A. (2004). Blinded by science: The managerial consequences of inadequately validated cluster analysis solutions. *International Journal of Market Research*, *46*(2), 171–187.

In this article, the authors investigate if managers could distinguish between cluster analysis outputs derived from real-world and random data. They show that some managers feel able to assign meaning to random data devoid of a meaningful structure, and even feel confident formulating entire marketing strategies from cluster analysis solutions generated from such data. As such, the authors provide a reminder of the importance of validating clustering solutions with caution.

Dolnicar, S., Grün, B., & Leisch, F. (2016). Increasing sample size compensates for data problems in segmentation studies. *Journal of Business Research*, *69*(2), 992–999.

Using artificial datasets of known structure, the authors examine the effects of data problems such as respondent fatigue, sampling error, and redundant items on segment recovery. The study nicely shows how insufficient sample size of the segmentation base can have serious negative consequences on segment recovery and that increasing the sample size represents a simple measure to compensate for the detrimental effects caused by poor data quality.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2),134–148.

In this seminal article, the authors discuss several issues in applications of cluster analysis and provide further theoretical discussion of the concepts and rules of thumb that we included in this chapter.

Romesburg, C. (2004). *Cluster analysis for researchers*. Morrisville: Lulu Press.

Charles Romesburg nicely illustrates the most frequently used methods of hierarchical cluster analysis for readers with limited backgrounds in mathematics and statistics.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston: Kluwer Academic.

This book is a clear, readable, and interesting presentation of applied market segmentation techniques. The authors explain the theoretical concepts of recent analysis techniques and provide sample applications. Probably the most comprehensive text in the market.

References

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

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods in marketing research* (pp. 160–189). Cambridge: Basil Blackwell & Mott, Ltd..

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Philadelphia: Society for Industrial and Applied Mathematics.

Becker, J.-M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. *Marketing Letters, 26*(4), 643–659.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods, 3*(1), 1–27.

Dolnicar, S. (2003). Using cluster analysis for market segmentation—typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research, 11*(2), 5–12.

Dolnicar, S., & Grün, B. (2009). Challenging "factor-cluster segmentation". *Journal of Travel Research, 47*(1), 63–71.

Dolnicar, S., & Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of Marketing Management, 25*(3–4), 357–373.

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, F. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research, 53*(3), 296–306.

Dolnicar, S., Grün, B., & Leisch, F. (2016). Increasing sample size compensates for data problems in segmentation studies. *Journal of Business Research, 69*(2), 992–999.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification*. Hoboken: Wiley.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Hoboken: Wiley.

Everitt, B. S., & Rabe-Hesketh, S. (2006). *Handbook of statistical analyses using Stata* (4th ed.). Boca Raton: Chapman & Hall/CRC.

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung*. Beltz: Weinheim.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857–871.

Halpin, B. (2016). *Cluster analysis stopping rules in Stata*. University of Limerick. Department of Sociology Working Paper Series, WP2016-01. http://ulsites.ul.ie/sociology/sites/default/files/wp2016-01.pdf

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. Hoboken: Wiley.

Kotler, P., & Keller, K. L. (2015). *Marketing management* (15th ed.). Upper Saddle River: Prentice Hall.

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

Milligan, G. W., & Cooper, M. (1988). A study of variable standardization. *Journal of Classification, 5*(2), 181–204.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36*(2), 3336–3341.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148.

Qiu, W., & Joe, H. (2009). Cluster generation: Random cluster generation (with specified degree of separation). R package version 1.2.7.

Sheppard, A. (1996). The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores. *Tourism Analysis, 1*(1), 49–57.

Tonks, D. G. (2009). Validity and the design of market segments. *Journal of Marketing Management, 25*(3/4), 341–356.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston: Kluwer Academic.

van der Kloot, W. A., Spaans, A. M. J., & Heinser, W. J. (2005). Instability of hierarchical cluster analysis due to input order of the data: The PermuCLUSTER solution. *Psychological Methods, 10*(4), 468–476.

Lilien, G. L., & Rangaswamy, A. (2004). *Marketing engineering. Computer-assisted marketing analysis and planning* (2nd ed.). Bloomington: Trafford Publishing.

John H. R., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing, 31*(2), 127–140