Cluster Analysis

9

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 SPSS clustering outputs.

Keywords

Agglomerative and divisive clustering • Chebychev distance • City-block distance • Clustering variables • Dendrogram • Distance matrix • Euclidean distance • Hierarchical and partitioning methods • Icicle diagram • k-means • Matching coefficients • Profiling clusters • Two-step clustering

Thaltegos (http://www.thaltegos.com) is a German management consulting company, focusing on analytical approaches for marketing, sales and after sales challenges in the automotive industry. Due to their industry experience, a major US car manufacturer commissioned Thaltegos to develop a segmentation concept of the European car market. Using cluster analysis, Thaltegos identified three distinct segments comprising light-weight compact cars, sports cars, and limousines. Using this segmentation concept, the automotive manufacturer can derive concrete steps on how to position their new electric car in the market.

M. Sarstedt and E. Mooi, *A Concise Guide to Market Research*,
Springer Texts in Business and Economics, DOI 10.1007/978-3-642-53965-7_9,
© Springer-Verlag Berlin Heidelberg 2014

9.1 Introduction

One of the most fundamental marketing activities is in *market segmentation*. As companies cannot connect with all their potential customers, they have to divide markets into groups (segments) of consumers, customers, or clients with similar needs and wants. Firms can then target each of these segments by positioning themselves in a unique segment (such as Ferrari in the high-end sports car market). While market researchers often form market segments based on practical grounds, industry practice and wisdom, cluster analysis allows segments to be formed that are based on data that are less dependent on subjectivity.

The segmentation of customers is a standard application of cluster analysis, but it can also be used in different contexts such as evaluating typical supermarket shopping paths (Larson et al. 2005) or deriving employers' branding strategies (Moroko and Uncles 2009).

9.2 Understanding Cluster Analysis

Cluster analysis is a convenient method for identifying homogenous groups of objects called *clusters*. Objects (or cases, observations) in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster.

Let's try to gain a basic understanding of the cluster analysis procedure 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. 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 7-point scale with higher values denoting a higher degree of price consciousness and brand loyalty. The values of seven respondents are shown in Table 9.1 and the scatter plot in Fig. 9.1.

The objective of cluster analysis is to identify groups of objects (in this case, customers) that are very similar with regard to their price consciousness and brand loyalty and assign them into clusters. After having decided on the clustering variables (brand loyalty and price consciousness), 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

Table 9.1 Data

Customer	A	B	C	D	E	F	G
x	3	6	5	3	6	4	1
y	7	7	6	5	5	3	2

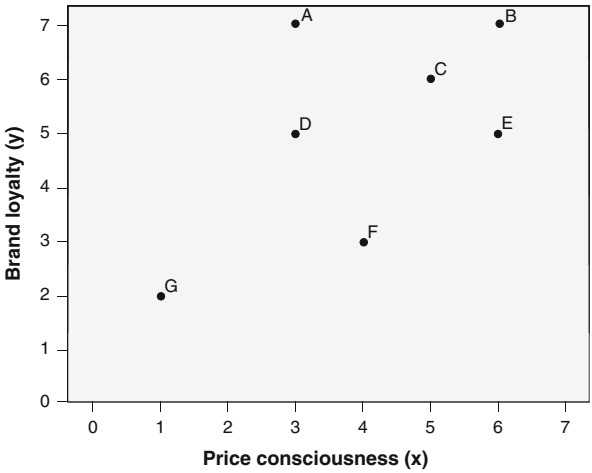


Fig. 9.1 Scatter plot

practice. We are going to discuss the most popular approaches in market research, as they can be easily computed using SPSS. These approaches are the following:

- *Hierarchical methods*,
- *Partitioning methods* (more precisely, *k-means*), and
- *Two-step clustering*.

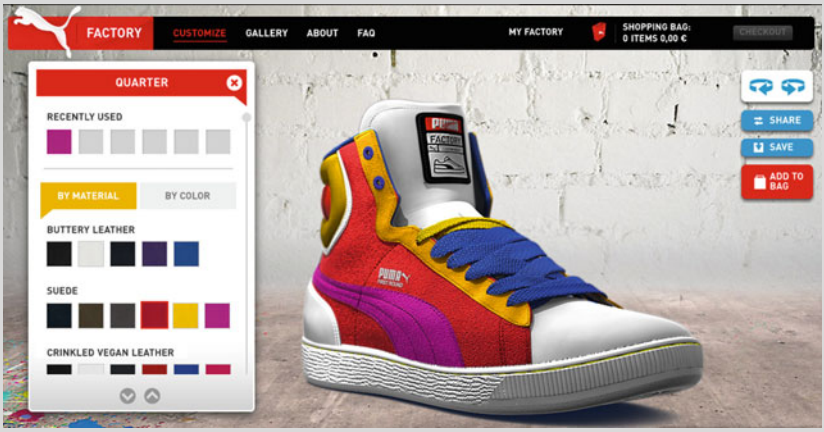
Each of these procedures follows a different approach to grouping the most similar objects into clusters. Specifically, whereas an object in a certain cluster should be as similar as possible to all the other objects in the same cluster, it should likewise be as distinct as possible from objects in different clusters.

But how do we measure similarity? Most methods calculate measures of (dis) similarity by estimating the distance between pairs of objects. Objects with smaller distances between one another are more similar, whereas objects with larger distances are more dissimilar.

An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. This question is explored in the next step of the analysis. Sometimes, we already know the number of segments that have to be derived from the data. For example, if we were asked to ascertain what characteristics distinguish frequent shoppers from infrequent ones, we need to find two different clusters. However, we do not usually know the exact number of clusters and then we face a trade-off. On the one hand, you want as few clusters as possible to make clusters easy to understand and actionable. On the other hand, having many clusters allows you to identify more segments and more subtle differences between segments. In an extreme case, you can address each individual separately (called *micromarketing*) to meet consumers' specific needs in the best possible way.

However, the costs associated with such a strategy may be prohibitively high in many business contexts. Thus, we have to ensure that the segments are large enough

Micromarketing in Practice



In the PUMA Factory, customers can fully customize a pair of shoes in a hands-on, tactile, and interactive shoe-making experience. This customization by the customers allows PUMA to target each customer individually with promotions or special offerings, allowing for a one-to-one interaction.

to make marketing programs profitable. As a result, we create some within-cluster heterogeneity, which makes targeted marketing programs less effective.

In the final step, we need to interpret the solution by defining and labeling the obtained clusters. This can be done by examining the clustering variables' mean values or by identifying explanatory variables to profile the clusters. Ultimately, managers should be able to identify customers in each segment 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 Decide on the Clustering Variables

At the beginning of the clustering process, we have to select appropriate variables for clustering. Even though this choice is of utmost importance, it is rarely treated as such and, instead, a mixture of intuition and data availability guide most analyses in marketing practice. However, faulty assumptions may lead to improper market segments and, consequently, to deficient marketing strategies. Thus, great care should be taken when selecting the clustering variables!

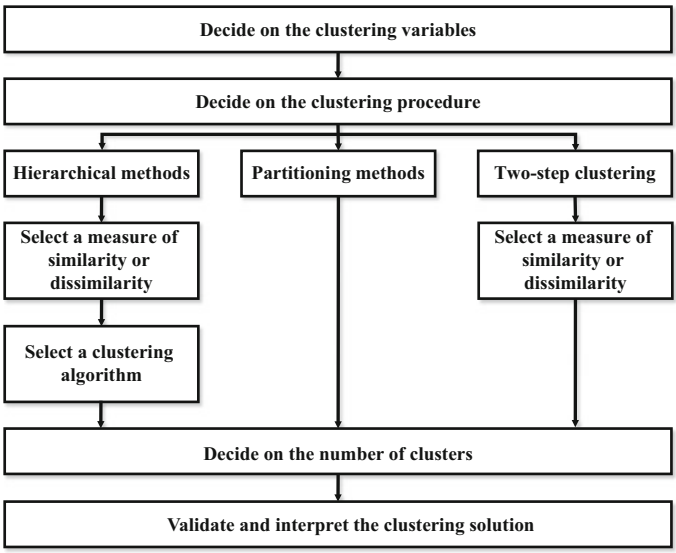


Fig. 9.2 Steps in a cluster analysis

There are several types of clustering variables and these can be classified as follows:

– *General* vs. *specific*, and
– *Observable* vs. *unobservable*.

General clustering variables are independent of products, services or circumstances whereas specific variables relate to both the customer and the product, service and/or particular circumstance. Furthermore, observable clustering variables can be measured directly while unobservable ones are inferred, for example, through observation or respondents' self-assessments. Table 9.2 provides several types and examples of clustering variables.

Table 9.2 Types and examples of clustering variables

	General	Specific
Observable	– Cultural – Demographic – Geographic – Socio-economic – ...	– Brand loyalty – Store loyalty – User status – Usage frequency – ...
Unobservable	– Lifestyle – Personality – Psychographics – Values – ...	– Attitudes – Intentions – Perceptions – Preferences – ...

The types of variables used for cluster analysis provide different segments and, thereby, influence segment-targeting strategies. Over the last decades, attention has shifted from more traditional general clustering variables towards product-specific unobservable variables. The latter generally provide better guidance for decisions on marketing instruments' effective specification. Generally, segments identified by means of specific unobservable variables are more homogenous and their consumers respond more consistent to marketing actions (see Wedel and Kamakura 2000). However, consumers in these segments are also frequently hard to identify from variables that are easily measured, such as demographics. Conversely, segments determined by means of observable variables usually stand out due to their identifiability but often lack a unique response structure.[1] Consequently, researchers frequently combine different variables (e.g., multiple lifestyle characteristics combined with 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, you should consider the following guiding questions:

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

Do the variables sufficiently differentiate the segments?

It is important to select those clustering variables that provide a clear-cut differentiation between the segments regarding a specific managerial objective.[2] More precisely, criterion validity is of special interest; that is, the extent to which the "independent" clustering variables are associated with one or more "dependent" variables not included in the analysis. Such "dependent" variables typically relate to some aspect of behavior, such as purchase intention or willingness-to-pay. Given this relationship, there should be significant differences between the "dependent" 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 the variable(s) of interest significantly.

[1] See Wedel and Kamakura (2000).

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

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, specific aspects covered by these variables 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 virtually cover the same aspect as *brand loyalty*. Thus, the concept of being attached to a brand would be overrepresented in the analysis because the clustering procedure does not differentiate between the clustering variables in a conceptual sense. Researchers frequently handle such correlation problems by applying cluster analysis to the observations' factor scores derived from a previously carried out factor analysis. However, this so called *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:

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 a certain amount of variance; thus, information is discarded before segments 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 segments are discarded, making it impossible to ever identify such groups.
4. The interpretations of clusters based on the original variables become questionable given that the segments have been constructed using factor scores.

Several studies have shown that the factor-cluster segmentation significantly reduces the success of finding useable segments.[3] Consequently, you should rather reduce the number of items in the questionnaire's pre-testing phase, retaining a reasonable number of relevant, non-redundant questions that you believe differentiate the segments well. However, if you have your doubts about the data structure, factor-clustering segmentation may still be a better option than discarding items that may conceptually be necessary.

[3] See the studies by Arabie and Hubert (1994), Sheppard (1996), or Dolnicar and Grün (2009).

Is the relation between sample size and number of clustering variables reasonable?
When choosing clustering variables, sample size is a point of concern. First and foremost, this relates to issues of managerial relevance as segment sizes need to be substantial to ensure that 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 rule of thumb regarding minimum sample sizes or the relationship between the objects and the number of clustering variables used. Formann (1984) recommends a minimum sample size of 2^m, where m equals the number of clustering variables. This rule-of-thumb can only provide rough guidance; nevertheless, we should pay attention to the relationship between the sample size and the number of clustering variables. It does not, for example, appear logical to cluster ten objects using ten variables. Keep in mind that no matter how many variables are used and no matter how small the sample size, cluster analysis will always render a result!

Are the data underlying the clustering variables of high quality?
Ultimately, the choice of clustering variables always depends on contextual influences such as data availability or resources to acquire additional data. Market researchers often overlook the fact 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). Data are of high quality if...
- ... the questions asked 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. For example, sales may wish to have segments that they can send the same salespeople to. Consequently, we 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 Decide on the Clustering Procedure

By choosing a specific clustering procedure, we determine how clusters are to be formed. This always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables' overall variance of objects in a specific cluster), or maximizing the distance between the

objects or 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 multimodal, exhaustive versus non-exhaustive).[4] A practical distinction is the differentiation between *hierarchical* and *partitioning methods* (most notably the *k-means* procedure), which we will discuss in the next sections. We also introduce *two-step clustering*, which combines the principles of hierarchical and partitioning methods and which has recently gained increasing attention from market research practice.

9.3.2.1 Hierarchical Methods

Understanding Hierarchical Clustering

Hierarchical clustering procedures are characterized by the tree-like structure established in the course of the analysis. Most hierarchical techniques fall into a category called *agglomerative clustering*. In this category, clusters are consecutively formed from objects. Initially, this type of procedure starts with each object representing an individual cluster. These clusters are then sequentially merged according to their similarity. First, the two most similar clusters are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on. This allows a hierarchy of clusters to be established from the bottom up. In Fig. 9.3 (left-hand side), we show how agglomerative clustering assigns additional objects to clusters step-by-step.

A cluster hierarchy can also be generated top-down. In this *divisive clustering*, all objects are initially merged into a single cluster, which is then gradually split up. Figure 9.3 illustrates this concept (right-hand side). 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. This is an important distinction between these types of clustering and partitioning methods such as k-means, which we will explore in the next section.

Divisive procedures are quite rarely used in market research. We therefore concentrate on the agglomerative clustering procedures. There are various types of agglomerative procedures. However, before we discuss these, we need to define how similarities or dissimilarities are measured between pairs of objects.

[4] See Wedel and Kamakura (2000), Dolnicar (2003), and Kaufman and Rousseeuw (2005) for a review of clustering techniques.

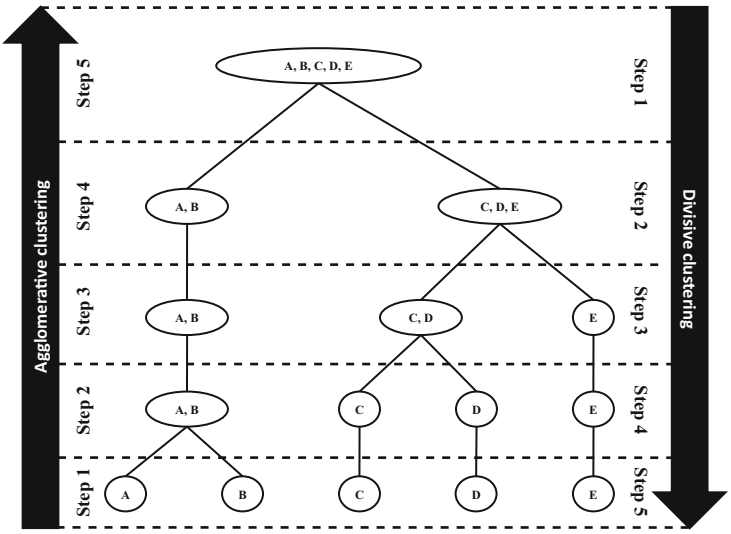


Fig. 9.3 Agglomerative and divisive clustering

Select a Measure of Similarity or Dissimilarity

Metric and Ordinal Variables

There are various measures to express (dis)similarity between pairs of objects. A straightforward way to assess two objects' proximity is by drawing a straight line between them. For example, when we look at 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 also referred to as *Euclidean distance* (or *straight-line distance*) and is the most commonly used type when it comes to analyzing ratio or interval-scaled variables.[5] In our example, we have ordinal variables, but market researchers usually treat ordinal variables as metric data to calculate distance metrics by assuming that the scale steps are equidistant (very much like in factor analysis, which we discussed in Chap. 8).

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)) using variables x and y by with the following formula:

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

[5] Note that researchers also often use the squared Euclidean distance.

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:

$$d_{Euclidean}(B,C) = \sqrt{(6-5)^2 + (7-6)^2} = \sqrt{2} = 1.414$$

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:

$$d_{Euclidean}(B,G) = \sqrt{(6-1)^2 + (7-2)^2} = \sqrt{50} = 7.071$$

Likewise, we can compute the distance between all other pairs of objects. All these distances are usually expressed by means of a *distance matrix*. In this distance matrix, the non-diagonal elements express the distances between pairs of objects and zeros on the diagonal (the distance from each object to itself is, of course, 0). In our example, the distance matrix is an 8 × 8 table with the lines and rows representing the objects (i.e., customers) under consideration (see Table 9.3). As the distance between objects B and C (in this case 1.414 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 loot at either the lower or upper non-diagonal elements.

An important feature of distance (and similarity) matrices are *ties*, which are, identical distances between two or more objects. For example, in Table 9.3, there are three pairs of objects with distances of 2.236. In fact, in the 21 cells, there are only 13 unique distance values. In practical applications (which usually rely on much more clustering variables and objects), ties are more the exception than the rule and generally don't have a pronounced impact on the results.

Table 9.3 Euclidean distance matrix

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

There are also alternative distance measures: The *city-block distance* uses the sum of the variables' absolute differences. This distance measure is often called 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:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |6-5| + |7-6| = 2$$

The resulting distance matrix is in Table 9.4.

Table 9.4 City-block distance matrix

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	3	2	0				
D	2	5	3	0			
E	5	2	2	3	0		
F	5	6	4	3	4	0	
G	7	10	8	5	8	4	0

Lastly, when working with metric (or ordinal) data, researchers frequently use the *Chebychev distance*, which is the maximum of the absolute difference in the clustering variables' values. In respect of customers B and C, this result is:

$$d_{Chebychec}(B,C) = \max(|x_B - x_C|, |y_B - y_C|) = \max(|6-5|, |7-6|) = 1$$

Figure 9.4 illustrates the interrelation between these three distance measures regarding two objects, C and G, from our example.

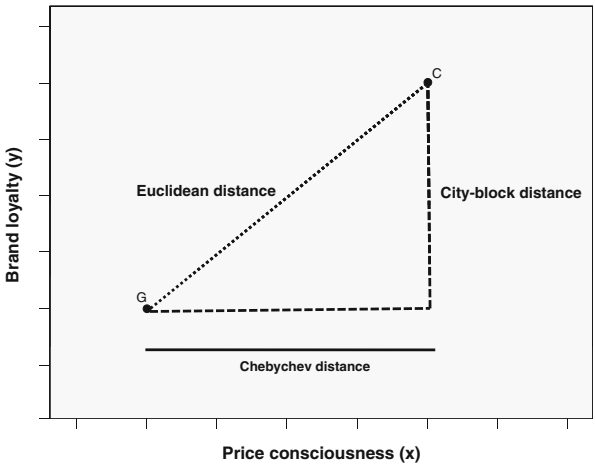


Fig. 9.4 Distance measures

There are other distance measures such as the *Angular*, *Canberra* or *Mahalanobis distances*. In many situations, the latter is desirable as it compensates for collinearity between the clustering variables. However, it is (unfortunately) not menu-accessible in SPSS.

In many analysis tasks, the variables under consideration have a different number of categories. This would be the case if we extended our set of clustering variables by adding another ordinal variable representing the customers' income measured by means of, for example, 15 categories. Since the absolute variation of the income variable would be much greater than the variation of the remaining two variables (remember, that x and y are measured on 7-point scales), this would clearly distort our analysis results. We can resolve this problem by standardizing the data prior to the analysis.

Different standardization methods are available, such as the simple z standardization, which rescales each variable to have a mean of 0 and a standard deviation of 1 (see Chap. 5). In most situations, however, standardization by range (e.g., to a range of 0 to 1 or −1 to 1) performs better.[6] We recommend standardizing the data in general, even though this procedure can reduce or inflate the variables' influence on the clustering solution.

Another way of (implicitly) standardizing the data is by using the correlation between the objects instead of distance measures. For example, suppose a respondent rated price consciousness 2 and brand loyalty 3. Now suppose a second respondent indicated 5 and 6, whereas a third rated these variables 3 and 3. Euclidean, city-block, and Chebychev distances would 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. Using correlation is also a way of standardizing the data implicitly.

Whether you use correlation or one of the distance measures 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 objects (which favors distance). However, it is generally recommended that one uses correlations when applying clustering procedures that are susceptible to outliers, such as complete linkage, average linkage or centroid (see section "Select a Clustering Algorithm").

[6] See Milligan and Cooper (1988).

Nominal Variables

Whereas the distance measures presented thus far can be used for variables measured on a metric and, in general, ordinal scale, applying them to nominal variables is problematic. When nominal variables are involved you should rather select a similarity measure expressing the degree to which variables' values share the same category. These so-called *matching coefficients* can take different forms but rely on the same allocation scheme shown in Table 9.5. 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 features present in one but not the other object. This scheme applies to binary variables, that is, those with two categories. For 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).

Table 9.5 Allocation scheme for matching coefficients

		Object 2	
		Presence of a characteristics (1)	Absence of a characteristic (0)
Object 1	Presence of a characteristic (1)	a	b
	Absence of a characteristic (0)	c	d

Based on the allocation scheme in Table 9.5, we can compute different matching coefficients, such as the *simple matching coefficient (SM)*:

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

This coefficient takes into account both, the joint presence and the joint absence of a characteristic (as indicated by cells *a* and *d* in Table 9.5). This feature makes the simple matching coefficient particularly useful for symmetric variables, where the joint presence and absence of a characteristic carries an equal degree of information (e.g., gender). However, if used on non-symmetric variables, objects may appear very similar because they both lack the same characteristics (as expressed through cell *d*) rather than because they share common characteristics (as expressed through cell *a*).

In light of this issue, researchers have proposed the *Jaccard (JC)* and the *Russel and Rao (RR)* coefficients, which do not include missing observations in the calculation of similarity (i.e., they (partially) omit the *d* cell from Table 9.5 in the calculation). Like

the simple matching coefficient, they range from 0 to 1 with higher values indicating a greater degree of similarity.[7] They are defined as follows:

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

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

To provide a brief example comparing the three coefficients, consider the following three variables:
- *Gender*: male, female
- *Product use*: light, medium, and heavy
- *Income*: low, medium, high

We consider the following two objects:
- *Object #1*: male customer, light user with low income and
- *Object #2*: female customer, medium user with a low income

We first transform the measurement data into binary data by recoding the original three variables into eight binary variables (i.e., two for gender and three for product use as well as income). The resulting binary data matrix is displayed in Table 9.6.

Table 9.6 Recoded measurement data

	Gender (binary code)		*Product use* (binary code)			*Income* (binary code)		
	Male	Female	Light	Medium	Heavy	Low	Medium	High
Object #1	1	0	1	0	0	1	0	0
Object #2	0	1	0	1	0	1	0	0

Using the allocation scheme from Table 9.5 yields the following results for the cells: a=1, b=2, c=2, and d=3.

This means that the two objects have only one shared characteristic (a=1), but three characteristics, which are absent from both objects (d=3). Using this information, we can now compute the three coefficients described earlier:

[7] There are many other matching coefficients such as *Yule's Q, Kulczynski* or *Ochiai*, but since most applications of cluster analysis rely on metric or ordinal data, we will not discuss these in greater detail. Check Wedel and Kamakura (2000) for more information on alternative matching coefficients.

$$SM = \frac{1+3}{1+2+2+3} = 0.5,$$

$$JC = \frac{1}{1+2+2} = 0.2, \text{and}$$

$$RR = \frac{1}{1+2+2+3} = 0.125$$

As can be seen, the simple matching coefficient suggests that the two objects are reasonably similar. On the contrary, the Jaccard coefficient and in particular the Russel Rao coefficient suggests that they are not.

Combinations of Metric, Ordinal, and Nominal Variables

Most datasets contain variables that are measured on multiple scales. For example, a market research questionnaire may ask about the respondent's income, product ratings, and last brand purchased. Thus, we have to consider variables measured on a ratio, ordinal, and nominal scale. How can we simultaneously incorporate these variables into one analysis? Unfortunately, this problem cannot be easily resolved. Often research use the distance measures discussed in the context of metric (and ordinal) data. Even though this approach may slightly change the results compared to using matching coefficients, it should not be rejected. Cluster analysis is mostly an exploratory technique whose results only provide a rough guidance for managerial decisions.

An alternative is to dichotomize all variables and apply the matching coefficients discussed above. For metric variables, this involves specifying categories (e.g., low, medium, and high income) 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 would lose precise information on each respondent's income, when scaling this variable down into income categories. In the light of these issues, you should avoid combining metric and nominal variables in a single cluster analysis. If this is not feasible, the *two-step clustering procedure* provides a valuable alternative, which we will discuss later. Lastly, the choice of the (dis)similarity measure is not very critical for the cluster structure. The choice of the clustering algorithm is far more important. We therefore deal with this first.

Select a Clustering Algorithm

After having chosen the distance or similarity measure, we need to decide which clustering algorithm to apply. There are several agglomerative procedures and they can be distinguished by the way they define the distance from a newly formed cluster to a certain object, or to other clusters in the solution. The most popular agglomerative clustering procedures 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.
– *Centroid*: 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.

Figures 9.5–9.8 illustrate these linkage procedures for two clusters.

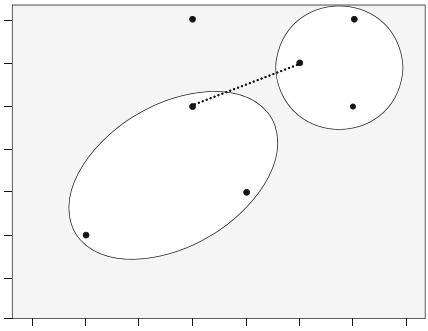


Fig. 9.5 Single linkage

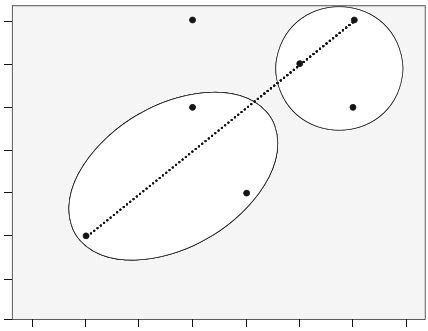


Fig. 9.6 Complete linkage

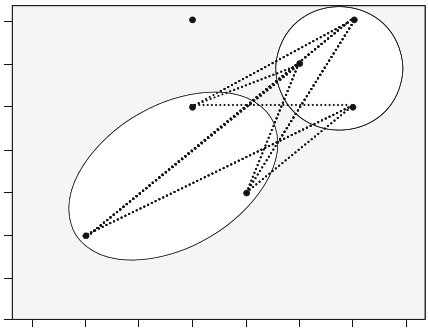


Fig. 9.7 Average linkage

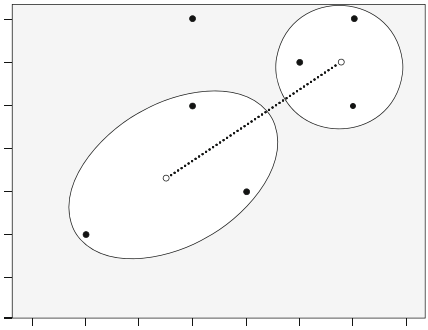


Fig. 9.8 Centroid

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its 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 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 rather compact and tightly clustered.
- The *average linkage* and *centroid* algorithms tend to produce clusters with low within-cluster variance and with similar sizes. Complete and average linkage are affected by outliers but less than the complete linkage method.

Another commonly used approach in hierarchical clustering is *Ward's method*. This approach does not combine the two most similar objects successively. Instead, those objects whose merger increases the overall within-cluster variance to the smallest possible degree, are combined. If you expect somewhat equally sized clusters and the dataset does not include outliers, you should always use Ward's method.

To better understand how a clustering algorithm works, let's manually examine some of the single linkage procedure's calculation steps. We start off by looking at the initial (Euclidean) distance matrix in Table 9.3. In the very first step, the two objects exhibiting the smallest distance in the matrix are merged. Note that we always merge those objects with the smallest distance, regardless of the clustering procedure (e.g., single or complete linkage). As we can see, this happens to two pairs of objects, namely B and C (d(B, C) = 1.414), as well as C and E (d(C, E) = 1.414). Depending on the clustering procedure used, this tie can lead to different clustering results. In this example we simply proceed by forming a new cluster using objects B and C.[8]

Having made this decision, we then form a new distance matrix by considering the single linkage decision rule as discussed above. According to this rule, the distance from, for example, object A to the newly formed cluster is the minimum of d(A, B) and d(A, C). As d(A, C)=2.236 is smaller than d(A, B)=3, the distance from A to the newly formed cluster is equal to d(A, C); that is, 2.236. We also compute the distances from cluster [B,C] (clusters are indicated by means of squared brackets) to all other objects (i.e. D, E, F, G) and simply copy the remaining distances—such as d(E, F)—that the previous clustering step has not affected. This yields the distance matrix shown in Table 9.7.

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

Objects	A	B, C	D	E	F	G
A	0					
B, C	2.236	0				
D	2	2.236	0			
E	3.606	1.414	3	0		
F	4.123	3.162	2.236	2.828	0	
G	5.385	5.657	3.606	5.831	3.162	0

[8] Note that because of ties, the final results may depend on the order of objects in the input file. Against this background, van der Kloot et al. (2005) recommend re-running the analysis with different input order of the data. At the same time, however, ties are more the exception than the rule in practical applications and generally don't have a pronounced impact on the results.

Continuing the clustering procedure, we simply repeat the last step by merging the objects in the new distance matrix that exhibit the smallest distance (in this case, the newly formed cluster [B, C] and object E) and calculate the distance from this new cluster to all other objects. The result of this step is described in Table 9.8.

Table 9.8 Distance matrix after second clustering step (single linkage)

Objects	A	B, C, E	D	F	G
A	0				
B, C, E	2.236	0			
D	2	2.236	0		
F	4.123	2.828	2.236	0	
G	5.385	5.657	3.606	3.162	0

Try to calculate the remaining steps yourself and compare your solution with the distance matrices in the following Tables 9.9–9.11.

Table 9.9 Distance matrix after third clustering step (single linkage)

Objects	A, D	B, C, E	F	G
A, D	0			
B, C, E	2.236	0		
F	2.236	2.828	0	
G	3.606	5.657	3.162	0

Table 9.10 Distance matrix after fourth clustering step (single linkage)

Objects	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	2.236	0	
G	3.606	3.162	0

Table 9.11 Distance matrix after fifth clustering step (single linkage)

Objects	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	3.162	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 3.162. 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 in the dataset.

A common way to visualize the cluster analysis's progress is by drawing a *dendrogram*, which displays the distance level at which there is a merger of objects and clusters (Fig. 9.9).

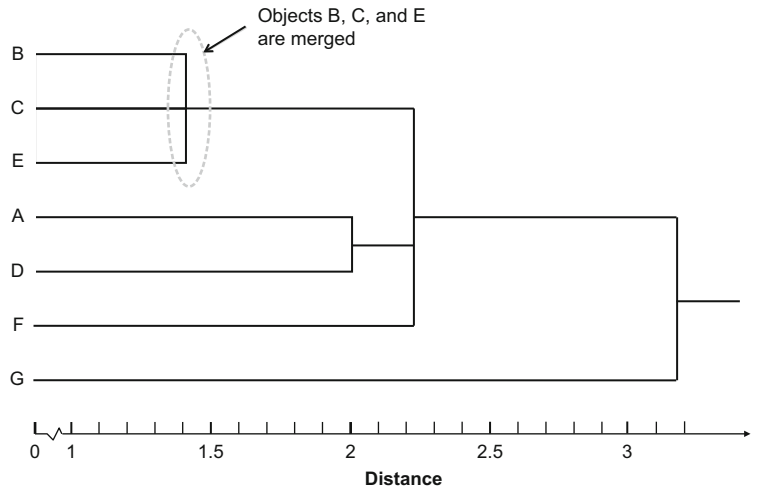


Fig. 9.9 Dendrogram

We read the dendrogram from left to right. The vertical lines indicate the distances at which objects have been combined. For example, according to our calculations above, objects B, C, and E are merged at a distance of 1.414.

Decide on the Number of Clusters

An important question we haven't yet addressed is how to decide on the number of clusters. Unfortunately, hierarchical methods provide only very limited guidance for making this decision. The only meaningful indicator relates to the distances at which the objects are combined. Similar to the scree plot in factor analysis, 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.

One potential way to solve this problem is to plot the number of clusters on the x-axis (starting with the one-cluster solution at the very left) against the distance at which objects or clusters are combined on the y-axis. Using this plot, we then search for the distinctive break (*elbow*).

Alternatively, we can make use of the dendrogram which essentially carries the same information. SPSS provides a dendrogram; however, it differs slightly from the one presented in Fig. 9.9 as SPSS rescales the distances to a range of 0–25 (i.e., the last merging step to a one-cluster solution takes place at a rescaled distance of 25). The rescaling often lengthens the merging steps, thus making breaks occurring at a greatly increased distance level more obvious.

However, this distance-based decision rule does not work very well in all cases. It is often difficult to identify where the break actually occurs. This is also the case in our example above. By looking at the dendrogram, we could justify a two-cluster solution ([A,B,C,D,E,F] and [G]), as well as a five-cluster solution ([B,C,E], [A], [D], [F], [G]).

Research has suggested several other procedures for determining the number of clusters in a dataset. Most notably, the *variance ratio criterion* (VRC) by Calinski

and Harabasz (1974) works well in many situations.[9] For a solution with n objects and k segments, the VRC is as follows:

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

where SS_B is the sum of the squares between the segments and SS_W is the sum of the squares within the segments. The criterion should seem familiar, as this is similar to the F-value of a one-way ANOVA. Consequently, the VRC can easily be computed using SPSS, even though it is not readily available in the clustering procedures' outputs. To finally determine the appropriate number of segments, we compute ω_k for each segment solution as follows:

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

In the next step, we choose the number of segments k that minimizes the value in ω_k. Owing to the term VRC_{k-1}, the minimum number of clusters that can be selected is three, which is a clear disadvantage of the criterion, thus limiting its application in practice.

Overall, the data can often only provide rough guidance regarding the number of clusters you should select; consequently, you should rather revert to practical considerations. 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.

9.3.2.2 Partitioning Methods: k-means

Understanding k-means Clustering

Another important group of clustering procedures are partitioning methods. As with hierarchical clustering, there is a wide array of different algorithms; of these, the *k-means procedure* is the most important one for market research.[10] The k-means algorithm follows an entirely different concept than the hierarchical methods discussed before. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the within-cluster variation as a

[9] Milligan and Cooper (1985) compare various criteria.

[10] Note that the k-means algorithm is one of the simplest non-hierarchical clustering methods. Several extensions, such as *k-medoids* (Kaufman and Rousseeuw 2005) have been proposed to handle limitations of the procedure. More advanced methods include finite mixture models (McLachlan and Peel 2000), neural networks (Bishop 2006), and self-organizing maps (Kohonen 1982). Andrews and Currim (2003) discuss the validity of some of these approaches.

measure to form homogenous clusters. Specifically, the procedure aims at partitioning the data in such a way that the within-cluster variation is minimized.

The clustering process starts by randomly assigning objects to a (pre-specified) number of clusters. The objects are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster.

With the hierarchical methods, an object remains in a cluster once it is assigned to it, but with k-means, cluster affiliations can change in the course of the clustering process. Consequently, k-means does not build a hierarchy as described before (Fig. 9.3), which is why the approach is also frequently labeled as non-hierarchical. Another important property of k-means clustering is that we have to pre-specify the number of clusters prior to running the analysis. We discuss this issue later in this chapter.

For a better understanding of the approach, let's take a look at how it works in practice. Figs. 9.10–9.13 illustrate the four steps of the k-means clustering process:

- **Step 1:** Using the pre-specified number of clusters as input, the algorithm randomly selects a center for each cluster. In our example, two cluster centers are randomly initiated, which CC1 (first cluster) and CC2 (second cluster) in Fig. 9.10 represent.[11]
- **Step 2:** Euclidean distances are computed from the cluster centers to every single 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.

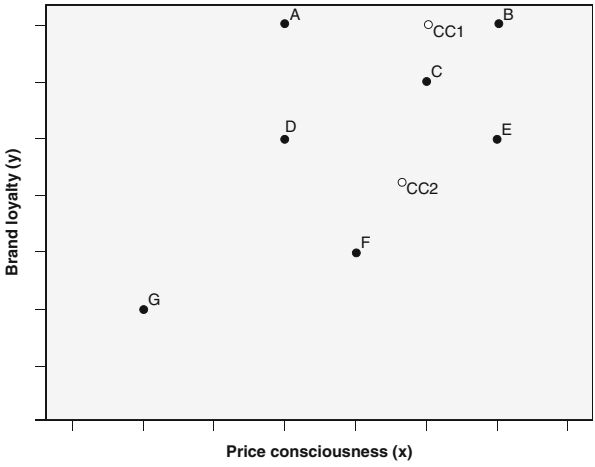


Fig. 9.10 k-means procedure (step 1)

[11] Conversely, SPSS always sets one observation as the cluster center instead of picking some random point in the dataset.

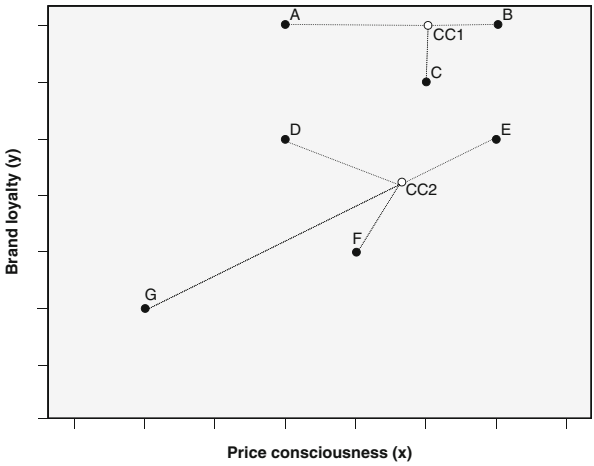


Fig. 9.11 k-means procedure (step 2)

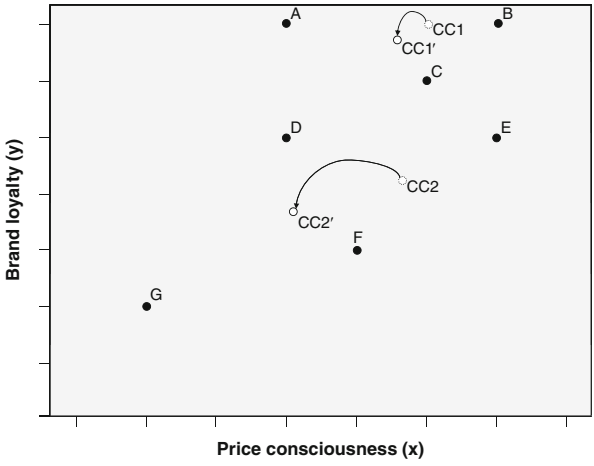


Fig. 9.12 k-means procedure (step 3)

- **Step 3:** Based on the initial partition from 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) regarding each of the variables (price consciousness and brand loyalty). As we can see in Fig. 9.12, both clusters' centers now shift into new positions (CC1' for the first and CC2' for the second cluster).
- **Step 4:** The distances from each object to the newly located cluster centers are computed and 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, as 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).

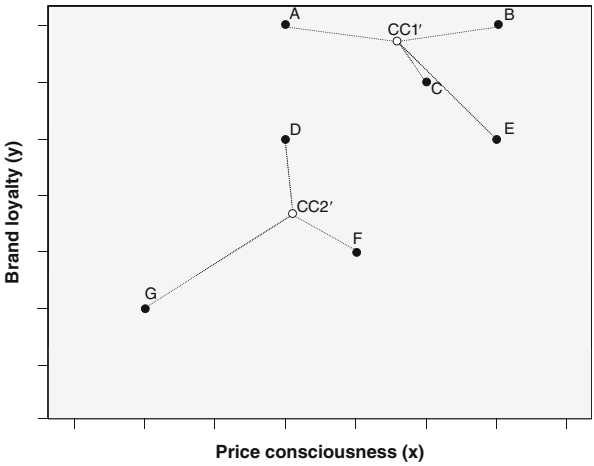


Fig. 9.13 k-means procedure (step 4)

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

Hierarchical or k-means clustering?
Generally, k-means is superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant clustering variables. Furthermore, k-means can be applied to very large datasets, as the procedure is less computationally demanding than hierarchical methods. In fact, we suggest k-means for sample sizes above 500, especially if many clustering variables are used. However, k-means should only be used on interval or ratio-scaled data as the procedure relies on Euclidean distances. Nevertheless, the procedure is routinely used on ordinal data as well, even though there might be some distortions. Finally, in k-means clustering, we have to pre-specify the number of clusters, which means we need to have some idea of the expected cluster solution before we start.

Decide on the Number of Clusters

When running k-means clustering, the researcher has to pre-specify the number of clusters to retain from the data. This makes k-means somewhat less attractive to some researchers and hinders its routine application in practice. Nevertheless, there are different ways to make this decision:

– Apply the VRC (discussed in the context of hierarchical clustering) on different number of clusters and chose the number that minimizes ω_k. See the 🖱 Web Appendix (→ Chap. 9) for an application of the VRC.

– Run a hierarchical procedure to determine the number of clusters and k-means afterwards.[12] 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.

Whatever approach you decide to choose, always keep in mind that cluster analysis is primarily an exploratory technique. Thus, practical considerations are of utmost importance when deciding on the number of clusters.

9.3.2.3 Two-Step Clustering

We have already discussed the issue of analyzing variables measured on different scale levels in this chapter. The *two-step cluster analysis* developed by Chiu et al. (2001) has been specifically designed to handle this problem. Like k-means, the procedure can also effectively cope with very large datasets.

The name two-step clustering is already an indication that the algorithm is based on a two-stage approach: In the first stage, the algorithm undertakes a procedure that is very similar to the k-means algorithm. Based on these results, the procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous clusters. This is done by building a so-called *cluster feature tree* whose "leaves" represent distinct objects in the dataset.

The procedure can handle categorical and continuous variables simultaneously and offers the user the flexibility to specify the cluster numbers as well as the maximum number of clusters, or to allow the technique to automatically choose the number of clusters on the basis of statistical evaluation criteria. Likewise, the procedure guides the decision of how many clusters to retain from the data by calculating measures of fit such as *Akaike's Information Criterion (AIC)* or *Bayes Information Criterion (BIC)*. These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments. "Relative" means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value. Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate a better fit. SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion.

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

However, which criterion should we choose? AIC is well-known for overestimating the "correct" number of segments, while BIC has a slight tendency to underestimate this number. Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results. But what do we do if the two criteria indicate different numbers of clusters? In such a situation, we should evaluate each solution on practical grounds as well as in light of the solution's interpretability. Do not solely rely on the automatic model selection, especially when there is a combination of continuous and categorical variables, as this does not always work well. Examine the results very carefully!

Two-step clustering also offers an overall goodness-of-fit measure called *silhouette measure of cohesion and sepearation*. It is essentially based on the average distances between the objects and can vary between -1 and $+1$. Specifically, a silhouette measure of less than 0.20 indicates a poor solution quality, a measure between 0.20 and 0.50 a fair solution, whereas values of more than 0.50 indicate a good solution. Furthermore, the procedure indicates each variable's importance for the construction of a specific cluster.

These desirable features make the somewhat less popular two-step clustering a viable alternative to the traditional methods. You can find a more detailed discussion of the two-step clustering procedure in the Web Appendix (→ Chap. 9), but we will also apply this method to the subsequent example.

9.3.3 Validate and Interpret the Cluster 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 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. Thus, the identified clusters should substantially *differ* from each other and 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. This step 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 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 2011; Tonks 2009).

- *Substantial*: The segments are large and profitable enough to serve.
- *Reliability*: Only segments that are stable over time can provide the necessary grounds for a successful marketing strategy. If segments 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 segments can be effectively reached and served, which requires them to be characterized by means of observable variables.
- *Actionable*: Effective programs can be formulated to attract and serve the segments.
- *Parsimonious*: To be managerially meaningful, only a small set of substantial clusters should be identified.
- *Familiar*: To ensure management acceptance, the segments composition should be comprehensible.
- *Relevant*: Segments should be relevant in respect of the company's competencies and objectives.
- *Compatibility*: Segmentation results meet other managerial functions' requirements.

9.3.3.1 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, in hierarchical clustering, 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 set-up 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.

Another common approach is to split the dataset into two halves and to analyze each separately using the same settings (i.e., the same clustering variables, procedure, number of segments, etc.). You then compare the two solutions' cluster centroids using, for example, t-tests of an ANOVA (see Chap. 6). If these do not differ significantly, you can presume that the overall solution has a high degree of stability.

When using hierarchical clustering, it is also worthwhile changing the order of the objects in your dataset and re-running the analysis to check the results' stability. As discussed earlier, due to ties in the distance matrix, hierarchical clustering can suffer from a non-uniqueness problem. If changing the order of the objects drastically changes the segment compositions (e.g., in terms of segment sizes), you should reconsider the set-up of the analysis and, for example, re-run it with different clustering variables.

9.3.3.2 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 segments 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 easily be 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.

9.3.3.3 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 segment 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 segment. For example, assume that we used a set of items to assess the respondents' values and learned that a certain segment comprises respondents who appreciate self-fulfillment, enjoyment of life, and a sense of accomplishment, whereas this is not the case in another segment. If we were able to identify explanatory variables such as gender or age, which distinguish these segments adequately, then we could assign a new person to a specific segment on the basis of these observable variables whose value traits may still be unknown.

9.3.3.4 Interpret the Clustering Solution

The interpretation of the solution requires characterizing each segment 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 segment have a lower willingness to pay and are satisfied with lower service levels, whereas customers in another segment are willing to pay more for a superior service. By using this information, we can also try to come up with 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 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 66 distinct segments to help marketers discern these consumers' likes, dislikes, lifestyles, and purchase behaviors.

An example is segment #51, called "Shotguns & Pickups," which comprises lower to middle-class families in rural areas in the US with a low to mid-level income (http://www.MyBestSegments.com).

Table 9.12 summarizes the steps involved in a hierarchical and k-means clustering. We also describe steps related to two-step clustering, which we will further introduce in the subsequent example.

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

Theory	Action
Research problem	
Identification of homogenous groups of objects in a population	
Select clustering variables that should be used 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 objects and clustering variables is reasonable (rule of thumb: Number of observations should be at least 2^m, where m is the number of clustering variables). Ensure that the sample size is large enough to guarantee substantial segments.
Low levels of collinearity among the variables	▶ Analyze ▶ Correlate ▶ Bivariate Eliminate or replace highly correlated variables (correlation coefficients > 0.90).

(continued)

Table 9.12 (continued)

Theory	Action
Specification	
Choose the clustering procedure	If there is a limited number of objects in your dataset or you do not know the number of clusters: ▶ Analyze ▶ Classify ▶ Hierarchical Cluster
	If there are many observations (> 500) in your dataset and you have a priori knowledge regarding the number of clusters: ▶ Analyze ▶ Classify ▶ K-Means Cluster
	If there are many observations in your dataset and the clustering variables are measured on different scale levels: ▶ Analyze ▶ Classify ▶ Two-Step Cluster
Select a measure of similarity or dissimilarity (only hierarchical and two-step clustering)	*Hierarchical methods*: ▶ Analyze ▶ Classify ▶ Hierarchical Cluster ▶ Method ▶ Measure
	Depending on the scale level, select the measure; convert variables with multiple categories into a set of binary variables and use matching coefficients; standardize variables if necessary (on a range of 0 to 1 or −1 to 1).
	Two-step clustering: ▶ Analyze ▶ Classify ▶ Two-Step Cluster ▶ Distance Measure
	Use Euclidean distances when all variables are continuous; for mixed variables, use log-likelihood.
Choose clustering algorithm (only hierarchical clustering)	▶ Analyze ▶ Classify ▶ Hierarchical Cluster ▶ Method ▶ Cluster Method
	Use Ward's method if equally sized clusters are expected and no outliers are present. Preferably use single linkage, also to detect outliers.
Decide on the number of clusters	*Hierarchical clustering:* Examine the dendrogram: ▶ Analyze ▶ Classify ▶ Hierarchical Cluster ▶ Plots ▶ Dendrogram
	Draw a scree plot (e.g., using Microsoft Excel) based on the coefficients in the agglomeration schedule.
	Compute the VRC using the ANOVA procedure: ▶ Analyze ▶ Compare Means ▶ One-Way ANOVA
	Move the cluster membership variable in the **Factor** box and the clustering variables in the **Dependent List** box.
	Compute VRC for each segment solution and compare values.
	k-means: Run a hierarchical cluster analysis and decide on the number of segments based on a dendrogram or scree plot; use this information to run k-means with k clusters.

(continued)

Table 9.12 (continued)

Theory	Action
	Compute the VRC using the ANOVA procedure: ▶ Analyze ▶ Classify ▶ K-Means Cluster ▶ Options ▶ ANOVA table; Compute VRC for each segment solution and compare values.
	Two-step clustering: Specify the maximum number of clusters: ▶ Analyze ▶ Classify ▶ Two-Step Cluster ▶ Number of Clusters
	Run separate analyses using AIC and, alternatively, BIC as clustering criterion: ▶ Analyze ▶ Classify ▶ Two-Step Cluster ▶ Clustering Criterion
	Examine the auto-clustering output.
Validate and interpret the cluster solution	
Stability	Re-run the analysis using different clustering procedures, algorithms or distance measures.
	Split the datasets into two halves and compute the clustering variables' centroids; compare centroids for significant differences (e.g., independent-samples t-test or one-way ANOVA).
	Change the ordering of objects in the dataset (hierarchical clustering only).
Differentiation of the data	Compare the cluster centroids across the different clusters for significant differences. 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.
Interpretation of the cluster solution	Identify names or labels for each cluster and characterize each cluster by means of observable variables.

9.4 Example

Thaltegos (http://www.thaltegos.com) is a German management consulting company focusing on analytical approaches for marketing, sales, and after sales in the automotive industry. A major US car manufacturer commissioned Thaltegos to support the launch of an innovative electric car. To better position the car in the market, the manufacturer asked Thaltegos to provide transparency concerning the

European car market. In cooperation with a market research firm, Thaltegos gathered data from major automotive manufacturers to develop a segmentation concept. The database consists of the following vehicle characteristics, all of which have been measured on a ratio scale (variable names in parentheses):

- Engine displacement (*displacement*)
- Turning moment in Nm (*moment*)
- Horsepower (*horsepower*)
- Length in mm (*length*)
- Width in mm (*width*)
- Net weight in kg (*weight*)
- Trunk volume in liters (*trunk*)
- Maximum speed in km/h (*speed*)
- Acceleration 0–100 km/h in seconds (*acceleration*)

The pretest sample of 15, randomly taken, cars is shown in Fig. 9.14. In practice, clustering is done on much larger samples but we use a small sample size to illustrate the clustering process. Keep in mind that in this example, the ratio between the objects and clustering variables is much too small. The dataset used is *thaltegos.sav* (Web Appendix → Chap. 9).

	Name	displacement	moment	horsepower	length	width	weight	trunk	speed	acceleration
1	Kia Picanto 1.1 Start	1086	97	65	3535	1595	929	127	154	15.10
2	Suzuki Splash 1.0	996	90	65	3715	1680	1050	178	160	14.70
3	Renault Clio 1.2	1149	105	75	3986	1719	1155	288	167	13.40
4	Dacia Sandero 1.6	1598	128	87	4020	1746	1111	320	174	11.50
5	Fiat Grande Punto 1.4	1598	140	88	3986	1719	1215	288	177	11.90
6	Peugot 207 1.4	1360	133	88	4030	1748	1214	270	180	12.70
7	Renault Clio 1.6	1368	125	95	4030	1687	1135	275	178	11.40
8	Porsche Cayman	3386	340	295	4341	1801	1340	410	275	5.40
9	Nissan 350Z	3498	353	301	4315	1815	1610	235	250	5.80
10	Mercedes C 200 CDI	2148	270	136	4595	1770	1605	485	208	10.80
11	VWPassat Variant 2.0	1968	320	140	4774	1820	1596	588	201	10.50
12	Skoda Octavia 2.0	1968	320	140	4572	1769	1425	580	207	9.70
13	Mercedes E 280	2996	300	231	4852	1822	1660	540	250	7.30
14	Audi A6 2.4	2393	230	177	4916	1855	1525	546	231	8.90
15	BMW 525i	2497	250	218	4841	1846	1550	520	245	7.50

Fig. 9.14 Data

In the next step, we will run several different clustering procedures on the data. We first apply a hierarchical cluster analysis based on Euclidean distances, using the single linkage method. This will help us determine a suitable number of segments, which we will use as input for a subsequent k-means clustering. Finally, we will run a two-step cluster analysis using SPSS.

9.4.1 Pre-analysis: Collinearity Assessment

Before we start with the clustering process, we have to examine the variables for substantial collinearity. Just by looking at the variable set, we suspect that there are some highly correlated variables in our dataset. For example, we expect rather high correlations between *speed* and *acceleration*. To determine this, we run a bivariate correlation analysis by clicking ▶ Analyze ▶ Correlate ▶ Bivariate, which will open a dialog box similar to that in Fig. 9.15. Enter all variables into the **Variables** box and select the box **Pearson** (under **Correlation Coefficients**) because these are continuous variables.

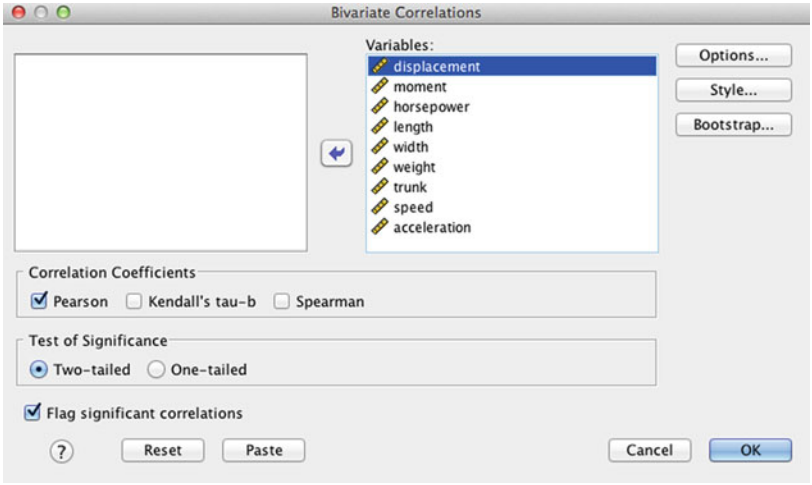


Fig. 9.15 Bivariate correlations dialog box

The correlation matrix in Table 9.13 supports our expectations – there are several variables that have high correlations. *Displacement* exhibits high (absolute) correlation coefficients with *horsepower*, *speed*, and *acceleration*, with values well above 0.90, indicating possible collinearity issues. Similarly, *horsepower* is highly correlated with *speed* and *acceleration*. Likewise, *length* shows a high degree of correlation with *width*, *weight*, and *trunk*.

A potential solution to this problem would be to run a factor analysis and perform a cluster analysis on the resulting factor scores. Since the factors obtained are, by definition, independent, this would allow for an effective handling of the collinearity issue. However, as this approach is associated with several problems (see Box 9.1), we should reduce the variables, for example, by omitting *displacement*, *horsepower*, and *length* from the subsequent analyses. The remaining variables still provide a sound basis for carrying out cluster analysis.

Table 9.13 Correlation matrix

Correlations

		displacement	moment	horsepower	length	width	weight	trunk	speed	acceleration
displacement	Pearson Correlation	1	.875**	.983**	.657**	.764**	.768**	.470	.967**	−.969**
	Sig. (2-tailed)		.000	.000	.008	.001	.001	.077	.000	.000
	N	15	15	15	15	15	15	15	15	15
moment	Pearson Correlation	.875**	1	.847**	.767**	.766**	.862**	.691**	.859**	−.861**
	Sig. (2-tailed)	.000		.000	.001	.001	.000	.004	.000	.000
	N	15	15	15	15	15	15	15	15	15
horsepower	Pearson Correlation	.983**	.847**	1	.608*	.732**	.714**	.408	.968**	−.961**
	Sig. (2-tailed)	.000	.000		.016	.002	.003	.131	.000	.000
	N	15	15	15	15	15	15	15	15	15
length	Pearson Correlation	.657**	.767**	.608*	1	.912**	.921**	.934**	.741**	−.714**
	Sig. (2-tailed)	.008	.001	.016		.000	.000	.000	.002	.003
	N	15	15	15	15	15	15	15	15	15
width	Pearson Correlation	.764**	.766**	.732**	.912**	1	.884**	.783**	.819**	−.818**
	Sig. (2-tailed)	.001	.001	.002	.000		.000	.001	.000	.000
	N	15	15	15	15	15	15	15	15	15
weight	Pearson Correlation	.768**	.862**	.714**	.921**	.884**	1	.785**	.778**	−.763**
	Sig. (2-tailed)	.001	.000	.003	.000	.000		.001	.001	.001
	N	15	15	15	15	15	15	15	15	15
trunk	Pearson Correlation	.470	.691**	.408	.934**	.783**	.785**	1	.579*	−.552*
	Sig. (2-tailed)	.077	.004	.131	.000	.001	.001		.024	.033
	N	15	15	15	15	15	15	15	15	15
speed	Pearson Correlation	.967**	.859**	.968**	.741**	.819**	.778**	.579*	1	−.971**
	Sig. (2-tailed)	.000	.000	.000	.002	.000	.001	.024		.000
	N	15	15	15	15	15	15	15	15	15
acceleration	Pearson Correlation	−.969**	−.861**	−.961**	−.714**	−.818**	−.763**	−.552*	−.971**	1
	Sig. (2-tailed)	.000	.000	.000	.003	.000	.001	.033	.000	
	N	15	15	15	15	15	15	15	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

9.4.2 Hierarchical Clustering

To run the hierarchical clustering procedure, click on ▶ Analyze ▶ Classify ▶ Hierarchical Cluster, which opens a dialog box similar to Fig. 9.16.

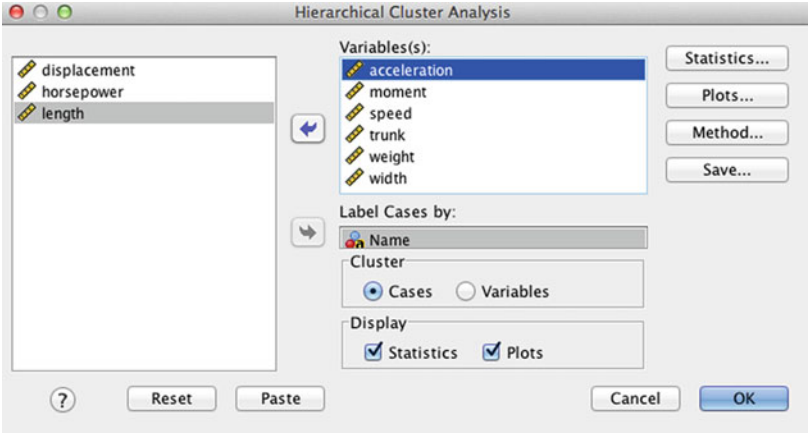


Fig. 9.16 Hierarchical cluster analysis dialog box

Move the variables *acceleration*, *moment*, *speed*, *trunk*, *weight*, and *width* into the **Variable(s)** box and specify *name* as the labeling variable (box **Label Cases by**). The **Statistics** option gives us the opportunity to request the distance matrix (labeled proximity matrix in this case) and the agglomeration schedule, which provides information on the objects being combined at each stage of the clustering process. Furthermore, we can specify the number or range of clusters to retain from the data. As we do not yet know how many clusters to retain, just check the box **Agglomeration schedule** and continue.

Under **Plots**, check the box **Dendrogram** to graphically display the distances at which objects and clusters are joined. Also ensure you select the **Icicle** diagram (**All clusters**), which is yet another graph for displaying clustering solutions.

The option **Method** allows us to specify the cluster method (e.g., single linkage or Ward's method), the distance measure (e.g., Chebychev distance or the Jaccard coefficient), and the type of standardization of values. In this example, we use the single linkage method (**Nearest neighbor**) based on **Euclidean distances**. Since the variables are measured on different levels (e.g., speed versus weight), make sure to standardize the variables, using the **Range −1 to 1** (**by variable**) in the **Transform Values** drop-down list.

Lastly, the **Save** option enables us to save cluster memberships for a single solution or a range of solutions. Saved variables can then be used in subsequent

analyses to explore differences between groups. As a start, we will skip this option, so continue and click on **OK** in the main menu.

First, we take a closer look at the agglomeration schedule (Table 9.14), which displays the objects or clusters combined at each stage (second and third column) and the distances at which this merger takes place. For example, in the first stage, objects 5 and 6 are merged at a distance of 0.149. From here onward, the resulting cluster is labeled as indicated by the first object involved in this merger, which is object 5. The last column on the very right tells you in which stage of the algorithm this cluster will appear next. In this case, this happens in the second step, where it is merged with object 7 at a distance of 0.184. The resulting cluster is still labeled 5, and so on. Similar information is provided by the *icicle diagram* shown in Fig. 9.17. Its name stems from the analogy to rows of icicles hanging from the eaves of a house. The diagram is read from the bottom to the top; the columns correspond to the objects being clustered, and the rows represent the number of clusters.

Table 9.14 Agglomeration schedule

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	6	.149	0	0	2
2	5	7	.184	1	0	3
3	4	5	.201	0	2	5
4	14	15	.213	0	0	6
5	3	4	.220	0	3	8
6	13	14	.267	0	4	11
7	11	12	.321	0	0	9
8	2	3	.353	0	5	10
9	10	11	.357	0	7	11
10	1	2	.389	0	8	14
11	10	13	.484	9	6	13
12	8	9	.575	0	0	13
13	8	10	.618	12	11	14
14	1	8	.910	10	13	0

As described earlier, we can use the agglomeration schedule to determine the number of segments to retain from the data. Next, we generate a *Scree Plot* by plotting the distances (**Coefficients** column in Table 9.14) against the number of clusters. The distinct break (elbow) indicates the solution regarding where an additional combination of two objects or clusters would occur at a greatly increased distance.

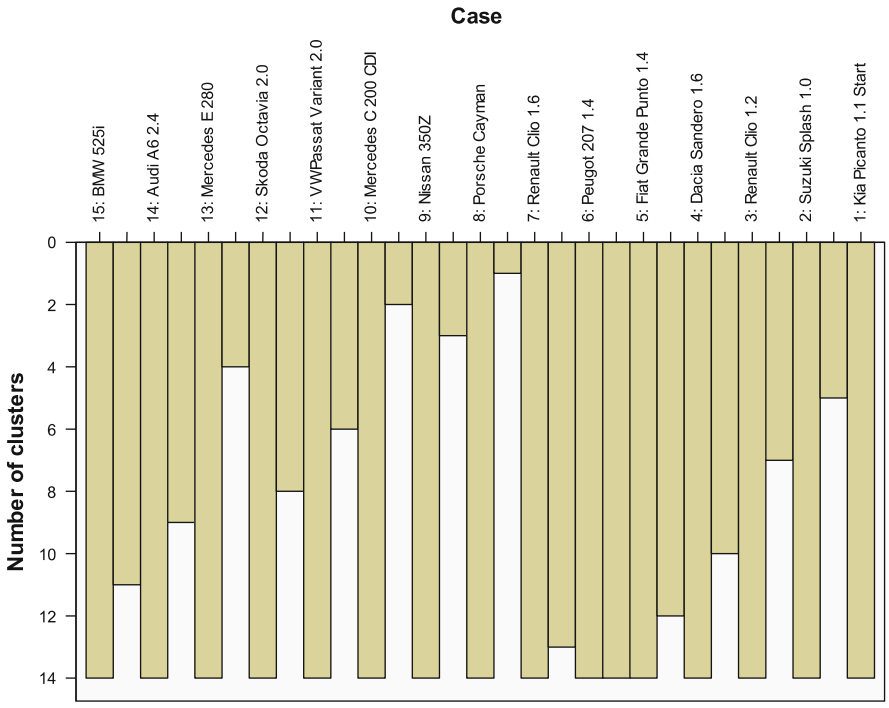


Fig. 9.17 Icicle diagram

Thus, the number of clusters prior to this merger is the most probable solution. SPSS does not automatically provide this plot. To generate a scree plot we have to double-click the **Agglomeration Schedule** in the output window. Next, we select all coefficients and right-click the mouse button. In the menu that opens up, we have to select **Create Graph ▶ Line** (Fig. 9.18). SPSS will add a line chart to the output which represents a scree plot. Note that in this plot, the *x*-axis represents the merging steps which means that, for example, the step from stage 13 to 14 represents the step from the two-cluster to the one-cluster solution. Note that – unlike in the factor analysis – we do not pick the solution with one cluster less than indicated by the elbow. The sharp increase in distance when switching from a one to a two-cluster solution occurs in almost all analyses and must not be viewed as a reliable indicator for the decision regarding the number of segments.

The scree plot in Fig. 9.19 shows that there is no clear elbow indicating a suitable number of clusters to retain. Based on the results, one could argue for a five-segment or six-segment solution. However, considering that there are merely 15 objects in the dataset, this seems too many, as we then have very small (and, most probably, meaningless) clusters. Consequently, a two, three or four-segment solution seems more appropriate.

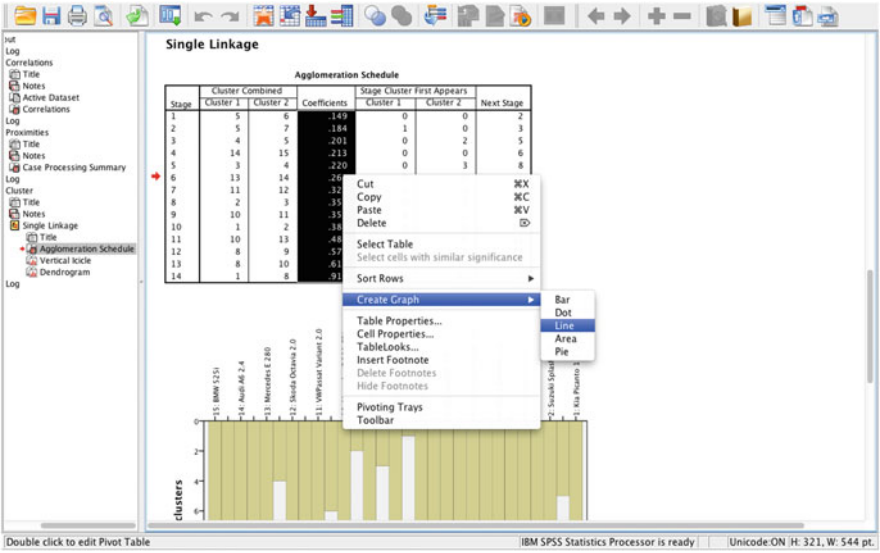


Fig. 9.18 Generating a scree plot

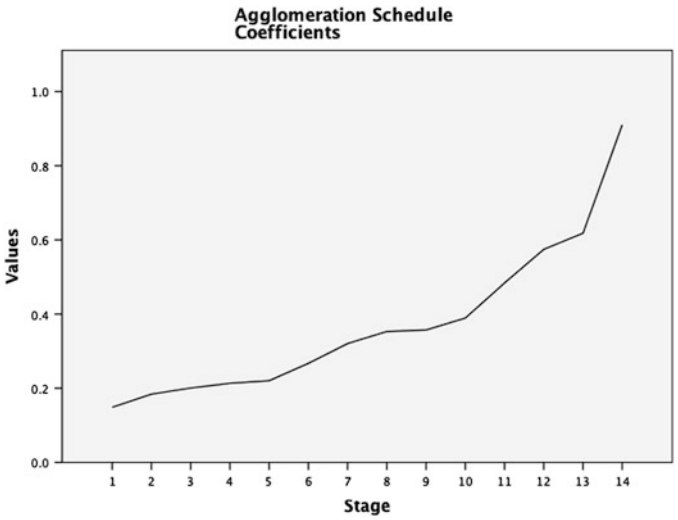


Fig. 9.19 Scree plot

Let's take a look at the dendrogram shown in Fig. 9.20. We read the dendrogram from the left to the right. Vertical lines are objects and clusters joined together – their position indicates the distance at which this merger takes place. When creating a dendrogram, SPSS rescales the distances to a range of 0–25; that is, the last merging step to a one-cluster solution takes place at a (rescaled) distance of 25. Note that this differs from our manual calculation shown in Fig. 9.9, where we did not do any rescaling! Again, the analysis only provides a rough guidance regarding the number of segments to retain. The change in distances between the mergers indicates that (besides a two-segment solution) both a three and four-segment solution are appropriate.

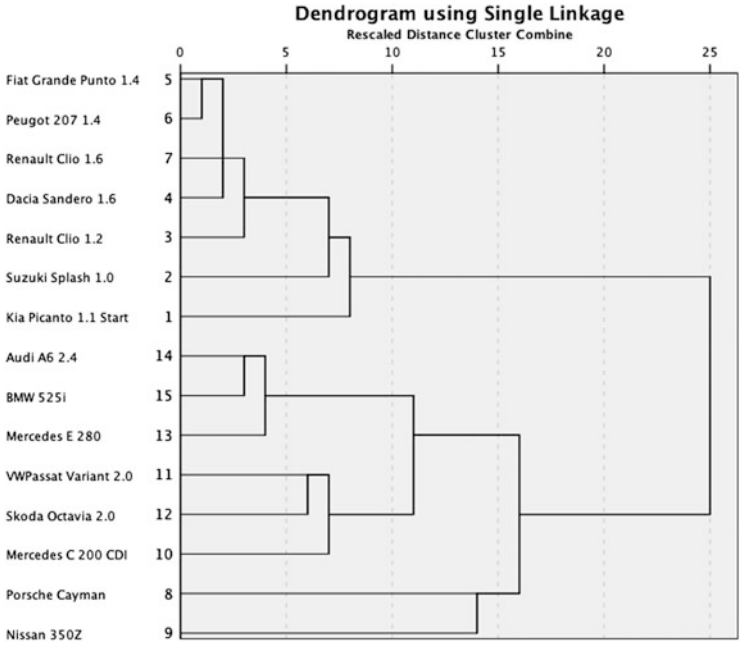


Fig. 9.20 Dendrogram

To clarify this issue, let's re-run the analysis, but this time we pre-specify different segment numbers to compare these with regard to content validity. To do so, just re-run the analysis using hierarchical clustering. Now switch to the **Save** option, specify a range of solutions from 2 to 4 and run the analysis. SPSS generates the same output but also adds three additional variables to your dataset (*CLU4_1*, *CLU3_1*, and *CLU2_1*), which reflect each object's cluster membership for the respective analysis. SPSS automatically places *CLU* in front, followed by the number of clusters (4, 3, or 2), to identify each object's cluster membership. Table 9.15 illustrates the results. SPSS does not produce this table for us, so we need to enter these cluster memberships ourselves in a table or spreadsheet.

Table 9.15 Cluster memberships

Name	Observation member of cluster (**four clusters**)	Observation member of cluster (**three clusters**)	Observation member of cluster (**two clusters**)
Kia Picanto 1.1 Start	1	1	1
Suzuki Splash 1.0	1	1	1
Renault Clio 1.2	1	1	1
Dacia Sandero 1.6	1	1	1
Fiat Grande Punto 1.4	1	1	1
Peugot 207 1.4	1	1	1
Renault Clio 1.6	1	1	1
Porsche Cayman	2	2	2
Nissan 350Z	3	2	2
Mercedes C 200 CDI	4	3	2
VW Passat Variant 2.0	4	3	2
Skoda Octavia 2.0	4	3	2
Mercedes E 280	4	3	2
Audi A6 2.4	4	3	2
BMW 525i	4	3	2

When we view the results, a three-segment solution appears promising. In this solution, the first segment comprises compact cars, whereas the second segment contains sports cars, and the third limousines. Increasing the solution by one segment would further split up the sports cars segment into two sub-segments. This does not appear to be very helpful, as now two of the four segments comprise only one object. This underlines the single linkage method's tendency to identify outlier objects—in this case the Nissan 350Z and Porsche Cayman. In this specific example, the Nissan 350Z and Porsche Cayman should not be regarded as outliers in a classical sense but rather as those cars which may be key competitors in the sports car market. In contrast, the two-segment solution appears to be rather imprecise considering the vast differences in the mix of sports and middle-sized cars in this solution.

To get a better overview of the results, let's examine the cluster centroids; that is, the mean values of the objects contained in the cluster on selected variables. To do so, we split up the dataset using the **Split File** command (▶ Data ▶ Split File) (see Chap. 5). This enables us to analyze the data on the basis of a grouping variable's values. In this case, we choose *CLU3_1* as the grouping variable and select the option **Compare groups**. Subsequently, we calculate descriptive statistics (▶ Analyze ▶ Descriptive Statistics ▶ Descriptives, also see Chap. 5) and calculate the mean, minimum and maximum values, as well as the standard deviations of the clustering variables. Table 9.16 shows the results for the variables *weight*, *speed*, and *acceleration*.

Table 9.16 Cluster centroids

Descriptive Statistics

CLU3_1		N	Minimum	Maximum	Mean	Std. Deviation
1	weight	7	929	1215	1115.57	100.528
	speed	7	154	180	170.00	9.950
	acceleration	7	11.40	15.10	12.9571	1.50317
	Valid N (listwise)	7				
2	weight	2	1340	1610	1475.00	190.919
	speed	2	250	275	262.50	17.678
	acceleration	2	5.40	5.80	5.6000	.28284
	Valid N (listwise)	2				
3	weight	6	1425	1660	1560.17	81.081
	speed	6	201	250	223.67	21.163
	acceleration	6	7.30	10.80	9.1167	1.48649
	Valid N (listwise)	6				

From the descriptive statistics, it seems that the first segment contains light-weight compact cars (with a lower maximum speed and acceleration). In contrast, the second segment comprises two sports cars with greater speed and acceleration, whereas the third segment contains limousines with an increased weight and intermediate speed and acceleration. Since the descriptives do not tell us if these differences are significant, we could use a one-way ANOVA (▶ Analyze ▶ Compare Means ▶ One-Way ANOVA) to calculate the cluster centroids and compare the differences formally.

9.4.3 k-means Clustering

In the next step, we want to use the *k-means* method on the data. We have previously seen that we need to specify the number of segments when conducting k-means clustering. SPSS then initiates cluster centers and assigns objects to the clusters based on their minimum distance to these centers. Instead of letting SPSS choose the centers, we can also save the centroids (cluster centers) from our previous analysis as input for the k-means procedure. To do this, we need to do some data management in SPSS, as the cluster centers have to be supplied in a specific format. Consequently, we need to aggregate the data first (briefly introduced in Chap. 5). By selecting ▶ Data ▶ Aggregate, a dialog box similar to Fig. 9.21 opens up. Note that we choose **Display Variable Names** instead of **Display Variable Labels** by clicking the right mouse button on the left box showing the variables in the dataset. Now we proceed by choosing the cluster membership variable (*CLU3_1*) as a break variable and move the *moment*, *width*, *weight*, *trunk*, *speed*, and *acceleration* variables into the **Summaries of Variable(s)** box. When using the default settings, SPSS computes the variables' mean values along the lines of the break variable (indicated by the postfix *_mean*, which is added to

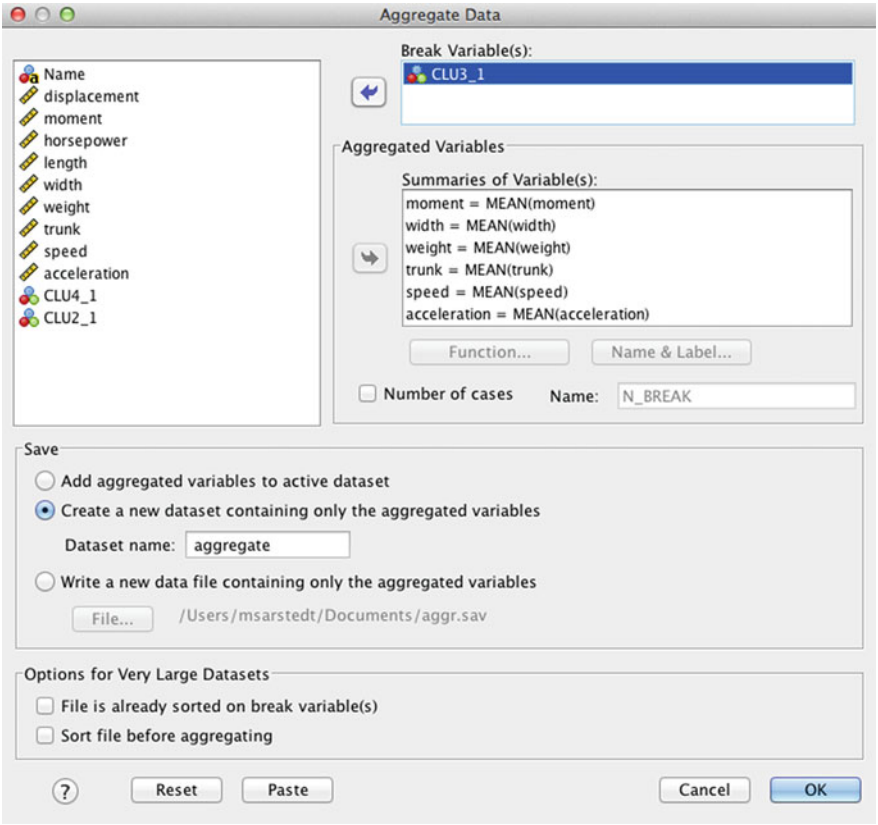


Fig. 9.21 Aggregate data dialog box

each aggregate variable's name), which corresponds to the cluster centers that we need for the k-means analysis. You can change each aggregate variable's name from the original one by removing the postfix *_mean* – using the **Name & Label** option – if you want to. Lastly, we do not want to add the aggregated variables to the active dataset, but rather need to create a new dataset comprising only the aggregated variables. You must therefore check this under **Save** and specify a dataset label such as *aggregate*. When clicking on **OK**, a new dataset labeled *aggregate* is created and opened automatically.

The new dataset is almost in the right format – but we still need to change the break variable's name from *CLU3_1* to *cluster_* (SPSS will issue a warning but this can be safely ignored). The final dataset should have the form shown in Fig. 9.22.

Now let's proceed by using k-means clustering. Make sure that you open the original dataset and go to Analyze ▶ Classify ▶ K-Means Cluster, which brings up a new dialog box (Fig. 9.23).

	cluster_	moment	width	weight	trunk	speed	acceleration
1	1	116.86	1699.14	1115.57	249.43	170.00	12.96
2	2	346.50	1808.00	1475.00	322.50	262.50	5.60
3	3	281.67	1813.67	1560.17	543.17	223.67	9.12

Fig. 9.22 Aggregated data file

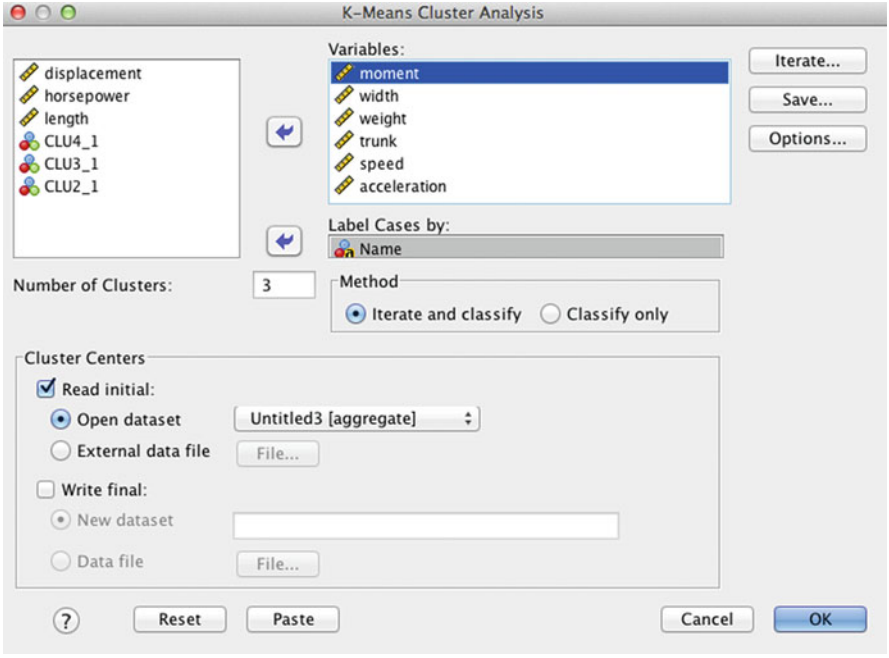


Fig. 9.23 K-means cluster analysis dialog box

As you did in the hierarchical clustering analysis, move the six clustering variables to the **Variables** box and specify the case labels (variable *name*). To use the cluster centers from our previous analysis, check the box **Read initial** and click on **Open dataset**. You can now choose the dataset labeled *aggregate*. Specify 3, which corresponds to the result of the hierarchical clustering analysis, in the **Number of Clusters** box. The **Iterate** option is of less interest to us. Instead, click on **Save** and check the box **Cluster Membership**. This creates a new variable indicating each object's final cluster membership. SPSS indicates whether each observation is a member of cluster 1, 2, or 3. Under **Options**, you can request several statistics and specify how missing values should be treated. Ensure that you request the **Initial cluster centers** as well as the **ANOVA table** and that you exclude the missing values listwise (default). Now start the analysis.

The k-means procedure generates Tables 9.17 and 9.18, which show the initial and final cluster centers. As you can see, these are identical (also compare Fig. 9.22), which indicates that the initial partitioning of the objects in the first step of the k-means procedure was retained during the analysis. This means that it was not possible to reduce the overall within-cluster variation by re-assigning objects to different clusters.

Table 9.17 Initial cluster centers

Initial Cluster Centers

	Cluster		
	1	2	3
moment	117	347	282
width	1699	1808	1814
weight	1116	1475	1560
trunk	249	323	543
speed	170	263	224
acceleration	12.96	5.60	9.12

Input from FILE Subcommand

Table 9.18 Final cluster centers

Final Cluster Centers

	Cluster		
	1	2	3
moment	117	347	282
width	1699	1808	1814
weight	1116	1475	1560
trunk	249	323	543
speed	170	263	224
acceleration	12.96	5.60	9.12

Likewise, the output **Iteration History** shows that there is no change in the cluster centers. Similarly, if you compare the partitioning of objects into the three clusters by examining the newly generated variable *QCL_1*, you see that there is no change in the clusters' composition. At first sight, this does not look like a very exciting result, but this in fact signals that the initial clustering solution is stable.

In other words, the fact that two different clustering methods yield the same outcomes provides some evidence of the results' stability.

In contrast to hierarchical clustering, the k-means outputs provide us with an ANOVA of the cluster centers (Table 9.19). As you can see, all the clustering variables' means differ significantly across at least two of the three segments, because the null hypothesis is rejected in every case (**Sig.** ≤ 0.05).

Table 9.19 ANOVA output

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
moment	64318.455	2	784.224	12	82.015	.000
width	23904.771	2	1966.183	12	12.158	.001
weight	339920.393	2	10829.712	12	31.388	.000
trunk	142764.143	2	4311.754	12	33.110	.000
speed	8628.283	2	262.153	12	32.913	.000
acceleration	50.855	2	2.057	12	24.722	.000

Since we used the prior analysis results from hierarchical clustering as an input for the k-means procedure, the problem of selecting the "correct" number of segments is not problematic in this example. As discussed above, we could have also used the VRC to make that decision. In the Web Appendix (→ Chap. 9), we present a VRC application to this example.

9.4.4 Two-step Clustering

As a last step of the analysis, we conduct a two-step clustering approach. First, go to Analyze ▶ Classify ▶ Two-Step Cluster. A new dialog box opens, similar to that shown in Fig. 9.24. First, move the variables we used in the previous analyses to the **Continuous Variables** box.

The **Distance Measure** box determines how the distance between two objects or clusters is computed. While **Log-likelihood** can be used for categorical and continuous variables, the **Euclidean** distance can only be applied when all of the variables are continuous. Unless your dataset contains categorical variables (e.g., gender) you should choose the Euclidean distance measure, as this generally provides better results. If you use ordinal variables and therefore use the **Log-likelihood** procedure, check that the answer categories are equidistant. In our dataset, all variables are continuous, therefore select the second option, namely **Euclidean**.

Under **Number of Clusters**, you can specify a fixed number or a maximum number of segments to retain from the data. One of two-step clustering's major

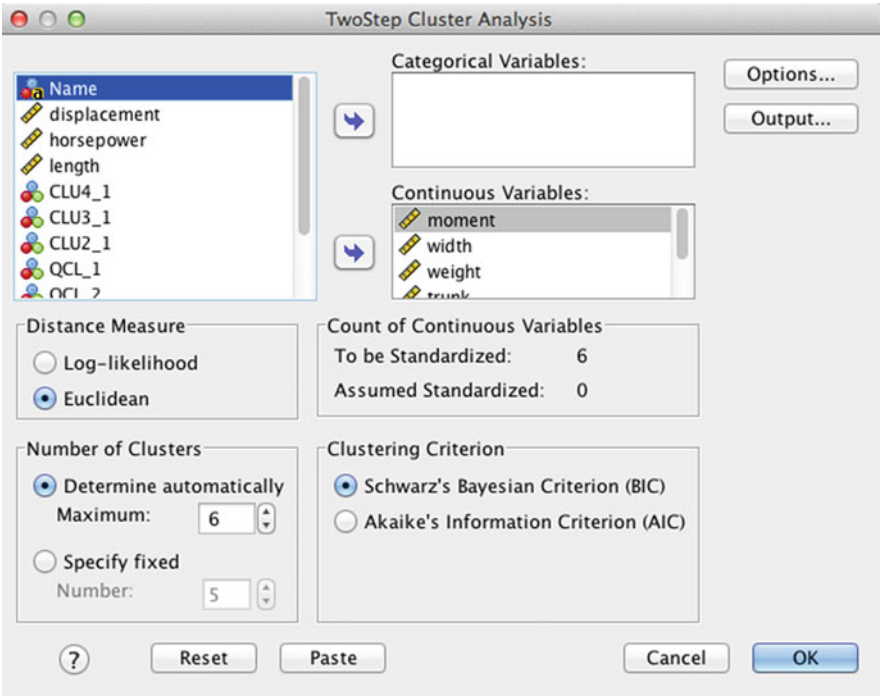


Fig. 9.24 Two-step cluster analysis dialog box

advantages is that it allows the automatic selection of the number of clusters. To make use of this advantage, you should specify a maximum number of clusters, for example, 6. Next to this box, in which the number of clusters is specified, you can choose between two criteria (also referred to as model selection or information criteria) which SPSS can use to pick an appropriate number of segments. Select Schwarz's Bayesian Criterion (BIC) but – as discussed above – you should re-run the analysis using AIC.

Under **Options**, you can specify issues related to outlier treatment, memory allocation, and variable standardization. Variables that are already standardized have to be assigned as such, but since this is not the case in our analysis, you can simply proceed.

Lastly, under the option **Output**, we can specify additional variables for describing the obtained clusters. However, let's stick to the default option for now. Make sure that you click the box **Create cluster membership variable** before clicking **Continue**.

SPSS produces a very simple output, as shown in Fig. 9.25. The upper part of the output describes the algorithm applied, the number of variables used (labeled input features) and the final number of clusters retained from the data. In our case, the number of clusters is chosen according to BIC, which indicates a two-segment solution (the same holds when using AIC instead of BIC). Note that this result differs from our previous analysis where we used a three-cluster solution!

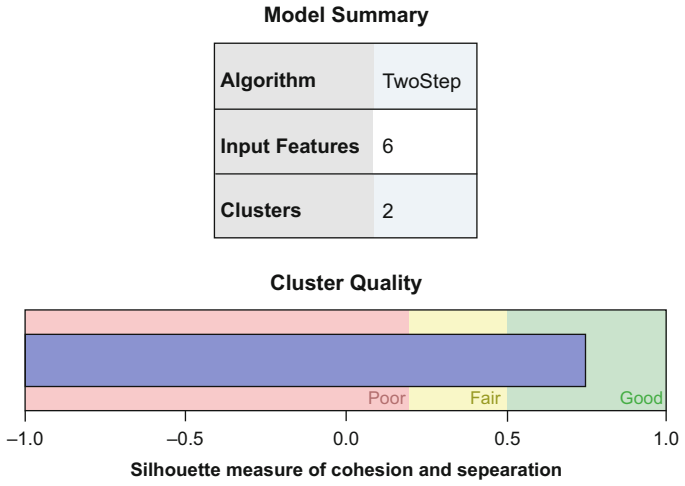


Fig. 9.25 Two-step clustering output

The lower part of the output (Fig. 9.25) indicates the quality of the cluster solution. The silhouette measure of cohesion and separation reaches a value of more than 0.50, indicating a satisfactory cluster quality. Consequently, you can proceed with the analysis by double-clicking on the output. This will open up the model viewer (Fig. 9.26), an evaluation tool that graphically presents the structure of the revealed clusters.

The model viewer provides us with two windows: The main view, which initially shows a model summary (left-hand side), and an auxiliary view, which initially features the cluster sizes (right-hand side). At the bottom of each window (option: **View**), you can request different information on each of the clusters. To further analyze the clusters, select **Clusters** in the main view and **Predictor Importance** in the auxiliary view (Fig. 9.26).

In the main view, we can now see a description of the two clusters, including their (relative) sizes. Furthermore, the output shows each clustering variables' mean values across the two clusters as well as their relative importance. Darker shades (i.e., higher values in feature importance) denote the variable's greater importance for the clustering solution (in terms of predicting each observation's cluster membership). Comparing the results, we can see that *moment* is the most important variable for each of the clusters, followed by *weight*, *speed*, *width*, *acceleration*, and *trunk*. Clicking on one of the boxes will show a graph with the frequency distribution of each cluster.

The auxiliary view on the right-hand side shows an overview of the variables' overall importance for predicting the clustering solution, which provides the same result as the cluster-specific analysis. The model viewer provides us with additional options for visualizing the results or comparing clustering solutions. It is

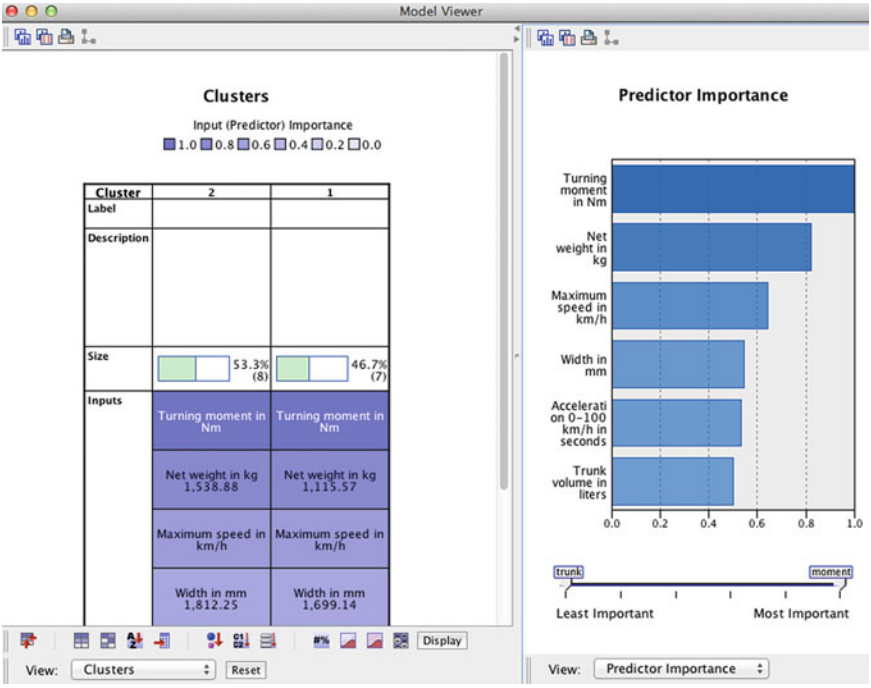


Fig. 9.26 Additional options in the model viewer

worthwhile to simply play around with the different self-explanatory options. So go ahead and explore the model viewer's features yourself!

9.5 Shopping at Projekt 2 (Case Study)

Facing dramatically declining sales and decreased turnover, retailers such as H&M and Zara are rethinking their pricing strategies, scaling back inventories, and improving the fashion content. Men's accessories are one of the bright spots and Projekt 2, an apparel retailer, has jumped on the trend with three recently opened shops prominently featuring this category. The largest men's store opened in Munich in 2011 and stocks top brands in jewelry, watches, sunglasses, and leather goods. By providing a better showcase for men's accessories, Projekt 2 aims at strengthening its position in a market that is often neglected in the larger department stores. This is because the men's accessories business generally requires expertise in buying since this typically involves small, artisan vendors – an investment many department stores are not willing to make.

Projekt 2's strategy seemed to be successful. However, before opening accessories shops in any other existing stores, the company wanted to gain further insights into their customers' preferences. Consequently, a survey was conducted

among visitors of the Munich store to gain a deeper understanding of their attitudes to buying and shopping. Overall, 180 respondents were interviewed using mall-intercept interviewing. The respondents were asked to indicate the importance of the following factors when buying products and services using a 5-point scale (1 = not at all important, 5 = very important):

- Saving time (x_1),
- Getting bargains (x_2),
- Getting products that aren't on the high street (x_3),
- Trying new things (x_4), and
- Being aware of what companies have to offer (x_5).

The resulting dataset *projekt2.sav* (Web Appendix → Chap. 9) also includes each respondent's gender and monthly disposable income.

1. Given the levels of measurement, which clustering method would you prefer? Carry out a cluster analysis using this procedure.
2. Interpret and profile the obtained clusters by examining cluster centroids. Compare differences across clusters on observed variables using ANOVA and post-hoc tests (see Chap. 6).
3. Use a different clustering method to test the stability of your results. If necessary, omit or rescale certain variables.
4. Based on your evaluation of the dataset, make recommendations to the management of Projekt 2's Munich store.

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. Run the k-means analysis again from the example application (*thaltegos.sav*, Web Appendix → Chap. 9). Compute a three-segment solution and compare the results with those obtained by the initial hierarchical clustering.
4. Run the k-means analysis again from the example application (*thaltegos.sav*, Web Appendix → Chap. 9). Use a factor analysis considering all nine variables and perform a cluster analysis on the resulting factor scores (factor-cluster segmentation). Interpret the results and compare these with the initial analysis.
5. Repeat the manual calculations of the hierarchical clustering procedure from the beginning of the chapter, but use complete or average linkage as clustering method. Compare the results with those of the single linkage method.
6. What clustering variables could be used to segment:
   - The market for smartphones?
   - The market for chocolate?
   - The market for car insurances?

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.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*, (4th edn). London: Arnold.

This book is comprehensive yet relatively non-mathematical, focusing on the practical aspects of cluster analysis. The authors discuss classical approaches as well as more recent methods such as finite mixture modeling and neural networks.

Journal of Classification. New York, NY: Springer, available at: http://www.springer.com/statistics/statistical+theory+and+methods/journal/357

If you are interested in the most recent advances in clustering techniques and have a strong background in statistics, you should check out this journal. Among the disciplines represented are statistics, psychology, biology, anthropology, archeology, astronomy, business, marketing, and linguistics.

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

Andrews, R. L., & Currim, I. S. (2003). Recovering and profiling the true segmentation structure in markets: An empirical investigation. *International Journal of Research in Marketing, 20*(2), 177–192.

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.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

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

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the 7th ACM SIGKDD international conference in knowledge discovery and data mining* (pp. 263–268). San Francisco, CA: Association for Computing Machinery.

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., & Grun, 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.

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

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

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59–69.

Kotler, P., & Keller, K. L. (2011). *Marketing management* (14th ed.). Upper Saddle River, NJ: Prentice Hall.

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing, 22*(4), 395–414.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

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.

Moroko, L., & Uncles, M. D. (2009). Employer branding and market segmentation. *Journal of Brand Management, 17*(3), 181–196.

Okazaki, S. (2006). What do we know about mobile internet adopters? A cluster analysis. *Information Management, 43*(2), 127–141.

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

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*, 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, NE: 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.

http://www.springer.com/978-3-642-53964-0

A Concise Guide to Market Research
The Process, Data, and Methods Using IBM SPSS Statistics
Sarstedt, M.; Mooi, E.
2014, XXII, 347 p. 119 illus., Hardcover
ISBN: 978-3-642-53964-0