Chapter 23
Cluster Analysis

Content list

By the end of this chapter you will understand:

1 The purposes of cluster analysis.
2 How to use SPSS to perform cluster analysis.
3 How to interpret the SPSS printout of cluster analysis.

Introduction

Cluster analysis is a major technique for classifying a 'mountain' of information into manageable meaningful piles. It is a data reduction tool that creates subgroups that are more manageable than individual datum. Like factor analysis, it examines the full complement of inter-relationships between variables. Both cluster analysis and discriminant analysis (Chapter 25) are concerned with classification. However, the latter requires prior knowledge of membership of each cluster in order to classify new cases. In cluster analysis there is no prior knowledge about which elements belong to which clusters. The grouping or clusters are defined through an analysis of the data. Subsequent multi-variate analyses can be performed on the clusters as groups.

Purpose of cluster analysis

Clustering occurs in almost every aspect of daily life. A factory's Health and Safety Committee may be regarded as a cluster of people. Supermarkets display items of similar nature, such as types of meat or vegetables in the same or nearby locations. Biologists have to organize the different species of animals before a meaningful description of the differences between animals is possible. In medicine, the clustering of symptoms and diseases leads to taxonomies of illnesses. In the field of business, clusters of consumer segments are often sought for successful marketing strategies.

Cluster analysis (CA) is an exploratory data analysis tool for organizing observed data (e.g. people, things, events, brands, companies) into meaningful taxonomies, groups, or clusters, based on combinations of IV's, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown. In this sense, CA creates new groupings without any preconceived notion of what clusters may arise, whereas discriminant analysis (Chapter 25) classifies people and items into already known groups. CA provides no explanation as to why the clusters exist nor is any interpretation made. Each cluster thus describes, in terms of the data collected, the class to which its members belong. Items in each cluster are similar in some ways to each other and dissimilar to those in other clusters.

A cluster. *A group of relatively homogeneous cases or observations.*

This visual example provides a clear picture of grouping possibilities.

	Jim	*Bev*	*Bob*	*Pat*	*Ryan*	*Carl*	*Tracy*	*Zac*	*Amy*	*Josh*
Score	*8*	*9*	*9*	*13*	*14*	*14*	*15*	*15*	*19*	*19*
		Cluster A				*Cluster B*			*Cluster C*	

It is fairly clear that, based on marks out of 20, there are three clusters. Jim, Bev and Bob form the lowest group, while Amy and Josh link as the top cluster. The remainder fall into a middle grouping. But with complex multivariate data, such eyeballing is not able to detect and assign persons or items to clusters as easily or as accurately as that. Cluster analysis is thus a tool of discovery revealing associations and structure in data which, though not previously evident, are sensible and useful when discovered. Importantly, it also enables new cases to be assigned to classes for identification and diagnostic purposes; or find exemplars to represent classes.

Imagine you wanted to undertake direct mail advertising with specific advertisements for different groups of people. You could use a variety of IV's like family income, age, number of cars per family, number of mobile phones per family, number of school children per family etc., to see if different postal or zip codes are characterized by particular combinations of demographic variables which could be grouped together to create a better way of directing the mail out. You might in fact find that postal codes could be grouped into a number of clusters, characterized as 'the retirement zone', 'nappy valley', 'the golf club set', the 'rottweiler in a pick-up' district, etc. This sort of grouping might also be valuable in deciding where to place several new wine stores, or 'Tummy to Toddler' shops.

Using cluster analysis, a customer 'type' can represent a homogeneous market segment. Identifying their particular needs in that market allows products to be designed with greater precision and direct appeal within the segment. Targeting specific segments is cheaper and more accurate than broad-scale marketing. Customers respond better to segment marketing which addresses their specific needs, leading to increased market share and customer retention. This is valuable, for example, in banking, insurance and tourism markets. Imagine four clusters or market segments in the vacation travel industry. They are: (1) The elite – they want top level service and expect to be pampered; (2) The escapists – they want to get away and just relax; (3) The educationalist – they want to see new things, go to museums, have a safari, or experience new cultures; (4) the sports person – they want the golf course, tennis court, surfing, deep-sea fishing, climbing, etc. Different brochures and advertising is required for each of these.

Brand image analysis, or defining product 'types' by customer perceptions, allows a company to see where its products are positioned in the market relative to those of its competitors. This type of modelling is valuable for branding new products or identifying possible gaps in the market. Clustering supermarket products by linked purchasing patterns can be used to plan store layouts, maximizing spontaneous purchasing opportunities.

Banking institutions have used hierarchical cluster analysis to develop a typology of customers, for two purposes, as follows:

- To retain the loyalty of members by designing the best possible new financial products to meet the needs of different groups (clusters), i.e. new product opportunities.
- To capture more market share by identifying which existing services are most profitable for which type of customer and improve market penetration.

One major bank completed a cluster analysis on a representative sample of its members, according to 16 variables chosen to reflect the characteristics of their financial transaction patterns. From this analysis, 30 types of members were identified. The results were useful for marketing, enabling the bank to focus on products which had the best financial performance; reduce direct mailing costs and increase response rates by targeting product promotions at those customer types most likely to respond; and consequently, to achieve better branding and customer retention. This facilitated a differential direct advertising of services and products to the various clusters that differed *inter alia* by age, income, risk taking levels, and self-perceived financial needs. In this way, the bank could retain and win the business of more profitable customers at lower costs.

Cluster analysis. *The statistical method of partitioning a sample into homogeneous classes to produce an operational classification.*

Cluster analysis, like factor analysis, makes no distinction between dependent and independent variables. The entire set of interdependent relationships are examined. Cluster analysis is the obverse of factor analysis. Whereas factor analysis reduces the number of variables by grouping them into a smaller set of factors, cluster analysis reduces the number of observations or cases by grouping them into a smaller set of clusters.

Qu. 23.1
For what purposes and situations would you choose to undertake cluster analysis? Check your answer against the information above.

The technique of cluster analysis

Because we usually don't know the number of groups or clusters that will emerge in our sample and because we want an optimum solution, a two-stage sequence of analysis occurs as follows:

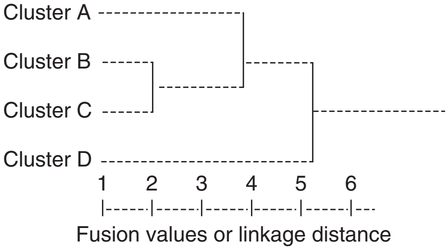
1 We carry out a *hierarchical cluster analysis* using *Ward's method* applying *squared Euclidean Distance* as the distance or similarity measure. This helps to determine the optimum number of clusters we should work with.
2 The next stage is to rerun the hierarchical cluster analysis with our selected number of clusters, which enables us to allocate every case in our sample to a particular cluster.

This sequence and methodology using SPSS will be described in more detail later. There are a variety of clustering procedures of which hierarchical cluster analysis is the major one.

Hierarchical cluster analysis

This is the major statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case as a separate cluster, i.e. there are as many clusters as cases, and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. The clustering method uses the dissimilarities or distances between objects when forming the clusters. The SPSS programme calculates 'distances' between data points in terms of the specified variables.

A hierarchical tree diagram, called a dendrogram on SPSS, can be produced to show the linkage points. The clusters are linked at increasing levels of dissimilarity. The actual measure of dissimilarity depends on the measure used, for example:



This example illustrates clusters B and C being combined at the fusion value of 2, and BC with A at 4. The fusion values or linkage distances are calculated by SPSS. The goal of the clustering algorithm is to join objects together into successively larger clusters, using some measure of similarity or distance. At the left of the dendrogram we begin with each object or case in a class by itself (in our example above there are only four cases). In very small steps, we 'relax' our criterion as to what is and is not unique. Put another way, we lower our threshold regarding the decision when to declare two or more objects to be members of the same cluster.

As a result, we *link* more and more objects together and *amalgamate* larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together as one cluster. In these plots, the horizontal axis denotes the fusion or linkage distance. For each node in the graph (where a new cluster is formed) we can read off the criterion distance at which the respective elements were linked together into a new single cluster. As a general process, clustering can be summarized as follows:

- The distance is calculated between all initial clusters. In most analyses, initial clusters will be made up of individual cases.
- Then the two most similar clusters are fused and distances recalculated.
- Step 2 is repeated until all cases are eventually in one cluster.

Qu. 23.2
Explain in your own words briefly what hierarchical cluster analysis does.
Check your answer against the material above.

Distance measures

Distance can be measured in a variety of ways. There are distances that are Euclidean (can be measured with a 'ruler') and there are other distances based on similarity. For example, in terms of kilometre distance (a Euclidean distance) Perth, Australia is closer to Jakarta, Indonesia, than it is to Sydney, Australia. However, if distance is measured in terms of the cities' characteristics, Perth is closer to Sydney (e.g. both on a big river estuary, straddling both sides of the river, with surfing beaches, and both English speaking, etc). A number of distance measures are available within SPSS. The squared Euclidean distance is the most used one.

Squared Euclidean distance

The most straightforward and generally accepted way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances, an extension of Pythagoras' theorem. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e. as if measured with a ruler). In a univariate example, the Euclidean distance between two values is the arithmetic difference, i.e. value1 – value2. In the bivariate case, the minimum distance is the hypotenuse of a triangle formed from the points, as in Pythagoras' theory. Although difficult to visualize, an extension of the Pythagoras' theorem will also give the distance between two points in n-dimensional space. The squared Euclidean distance is used more often than the simple Euclidean distance in order to place progressively greater weight on objects that are further apart. Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data.

Having selected how we will measure distance, we must now choose the clustering algorithm, i.e. the rules that govern between which points distances are measured to determine cluster membership. There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data. This is important since it tells us that, although cluster analysis may provide an objective method for the clustering of cases, there can be subjectivity in the choice of method. SPSS provides five clustering algorithms, the most commonly used one being Ward's method.

Ward's method

This method is distinct from other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In general, this method is very efficient. Cluster membership is assessed by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

k-means clustering

This method of clustering is very different from the hierarchical clustering and Ward method, which are applied when there is no prior knowledge of how many clusters there may be or what they are characterized by. K-means clustering is used when you already have hypotheses concerning the number of clusters in your cases or variables. You may want to 'tell' the computer to form exactly three clusters that are to be as distinct as possible. This is the type of research question that can be addressed by the *k*-means clustering algorithm. In general, the *k*-means method will produce the exact *k* different clusters demanded of greatest possible distinction.

Very frequently, both the hierarchical and the *k*-means techniques are used successively.

- The former (Ward's method) is used to get some sense of the possible number of clusters and the way they merge as seen from the dendrogram.
- Then the clustering is rerun with only a chosen optimum number in which to place all the cases (*k* means clustering).

One of the biggest problems with cluster analysis is identifying the *optimum number of clusters*. As the fusion process continues, increasingly dissimilar clusters must be fused, i.e. the classification becomes increasingly artificial. Deciding upon the optimum number of clusters is largely subjective, although looking at a dendrogram (see Fig. 23.1) may help. Clusters are interpreted solely in terms of the variables included in them. Clusters should also contain at least four elements. Once we drop to three or two elements it ceases to be meaningful.

Example

A keep fit gym group wants to determine the best grouping of their customers with regard to the type of fitness work programmes they want in order to facilitate timetabling and staffing by specially qualified staff. A hierarchical analysis is run and three major clusters stand out on the dendrogram between everyone being initially in a separate cluster and the final one cluster. This is then quantified using a *k*-means cluster analysis with three clusters, which reveals that the means of different measures of physical fitness measures do indeed produce the three clusters (i.e. customers in cluster 1 are high on measure 1, low on measure 2, etc.).

Interpretation of results

The cluster centroids produced by SPSS are essentially means of the cluster score for the elements of cluster. Then we usually examine the means for each cluster on each dimension using ANOVA to assess how distinct our clusters are. Ideally, we would obtain significantly different means for most, if not all dimensions, used in the analysis. The magnitude of the *F* values performed on each dimension is an indication of how well the respective dimension discriminates between clusters. It is useful to create on SPSS as you will see below a new variable on the data view file which indicates the cluster to which each case has been assigned. This cluster membership variable can be used in further analyses. Techniques for determining reliability and validity of clusters are as yet not developed. However, one could conduct cluster analysis using several different distance measures provided by SPSS and compare results. Alternatively, if the sample is large enough, it can be split in half with clustering performed on each and the results compared.

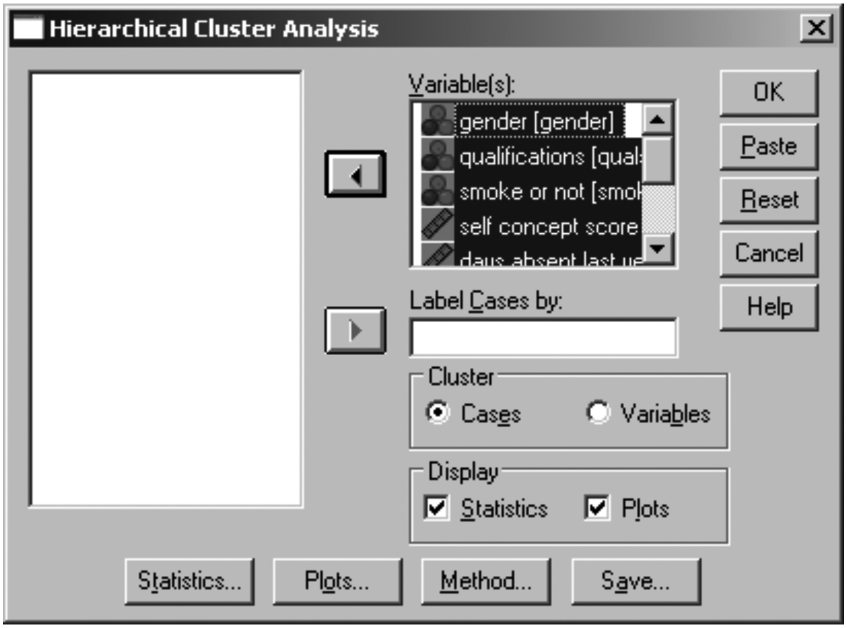
SPSS activity – conducting a cluster analysis

Access data file SPSS Chapter 23, Data File A on the website and load it on to your computer. This file only includes 20 cases, each responding to items on demographics (gender, qualifications, days absence from work, whether they smoke or not), on their attitudes to smoking in public places (subtest totals for pro and anti), plus total scale score for self-concept. We are attempting to determine how many natural groups exist and who belongs to each group.

The initial step is determining how many groups exist. The SPSS hierarchical analysis actually calculates every possibility between everyone forming their own group (as many clusters as there are cases) and everyone belonging to the same group, giving a range in our dummy set of data of from 1 to 20 clusters.

How to proceed

1 Click on ***Analyse*** > ***Classify*** > ***Hierarchical Cluster***.
2 Select variables for the analysis and transfer them to the ***Variables*** box (Fig. 23.1).
3 Click on ***Plots*** and select ***dendogram*** (Fig. 23.2).



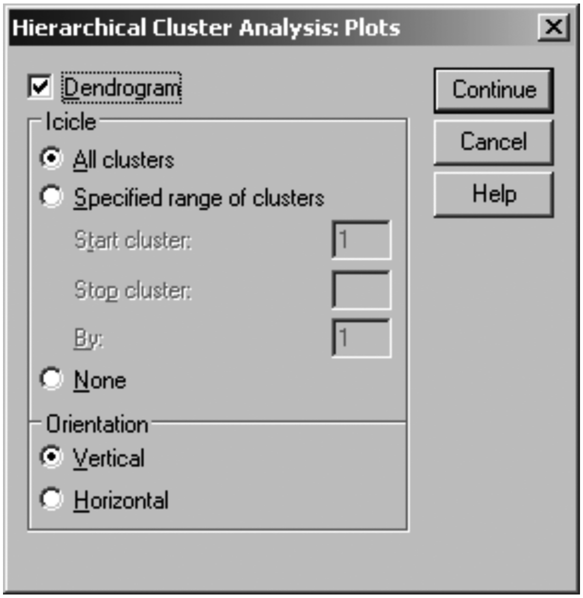
Variable(s):
gender [gender]
qualifications [quali
smoke or not [smok
self concept score
days absent last ye

Label Cases by:

Cluster
● Cases ○ Variables

Display
☑ Statistics ☑ Plots

Figure 23.1 **Hierarchical cluster analysis box.**



☑ Dendrogram

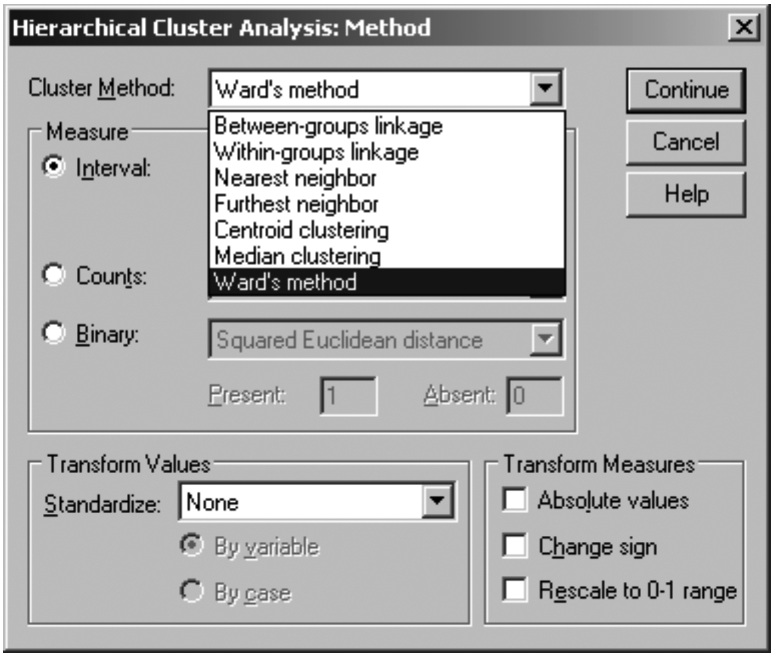
Icicle
● All clusters
○ Specified range of clusters
Start cluster: 1
Stop cluster:
By: 1
○ None

Orientation
● Vertical
○ Horizontal

Figure 23.2 **Hierarchical cluster analysis plots box.**

4 Click on ***Method*** and select *Ward's Method* (at the bottom of the Cluster Method list), and ensure ***Interval*** and ***Squared Euclidean Distance*** are selected (Fig. 23.3).

5 Select ***Continue*** then ***OK***.



Hierarchical Cluster Analysis: Method

Cluster Method: Ward's method

Measure

● Interval:

○ Counts:

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case

Transform Measures

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

Figure 23.3 Hierarchical cluster analysis method box.

Interpretation of the printout Table 23.1

The results start with an agglomeration schedule (Table 23.1) which provides a solution for every possible number of cluster from 1 to 20 (the number of our cases). The column to focus on is the central one which has the heading 'coefficients'. Reading from the bottom upwards, it shows that for one cluster we have an agglomeration coefficient of 3453.150, for two clusters 2362.438, for three clusters 1361.651, etc.

If we rewrite the coefficients as in Table 23.2 (not provided on SPSS) it is easier to see the changes in the coefficients as the number of clusters increase. The final column, headed 'Change', enables us to determine the optimum number of clusters. In this case it is 3 clusters as succeeding clustering adds very much less to distinguishing between cases.

The dendrogram may give support to the agglomeration schedule and in Figure 23.4 shows two clear clusters and a minor one.

6 Now we can rerun the hierarchical cluster analysis and request SPSS to place cases into one of three clusters. REPEAT STEPS 1 to 3 inclusive above.

Table 23.1 **Agglomeration schedule**

	Cluster combined			Stage cluster first appears		
Stage	**Cluster 1**	**Cluster 2**	**Coefficients**	**Cluster 1**	**Cluster 2**	**Next stage**
1	7	9	4.000	0	0	8
2	11	19	13.500	0	0	5
3	4	8	23.000	0	0	10
4	13	16	34.000	0	0	7
5	11	18	46.500	2	0	12
6	14	17	61.000	0	0	17
7	6	13	79.333	0	4	11
8	5	7	107.333	0	1	13
9	12	15	137.333	0	0	13
10	4	20	183.167	3	0	16
11	2	6	231.583	0	7	12
12	2	11	324.976	11	5	17
13	5	12	442.976	8	9	14
14	3	5	586.810	0	13	15
15	1	3	806.405	0	14	18
16	4	10	1055.071	10	0	19
17	2	14	1361.651	12	6	18
18	1	2	2362.438	15	17	19
19	1	4	3453.150	18	16	0

Table 23.2 **Re-formed agglomeration table**

No. of clusters	Agglomeration last step	Coefficients this step	Change
2	3453.150	2362.438	1090.712
3	2362.438	1361.651	1000.787
4	1361.651	1055.071	306.634
5	1055.071	806.071	248.946
6	806.405	586.810	219.595

A clear demarcation point seems to be here.

7 Click on ***Continue*** then on ***Save***. Select ***Single Solution*** and place ***3*** in the clusters box (Fig. 23.5). The number you place in the box is the number of clusters that seem best to represent the clustering solution in a parsimonious way. Finally click ***OK***.

A new variable has been generated at the end of your SPSS data file called clu3_1 (labelled Ward method in variable view). This provides the cluster membership for each case in your sample (Fig. 23.6).

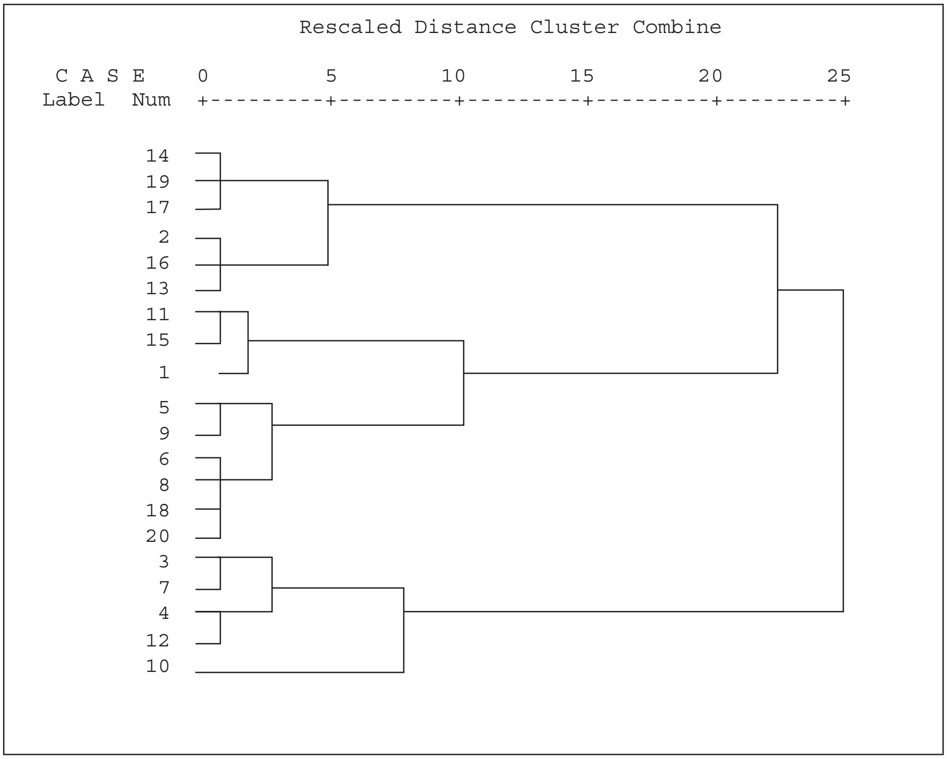
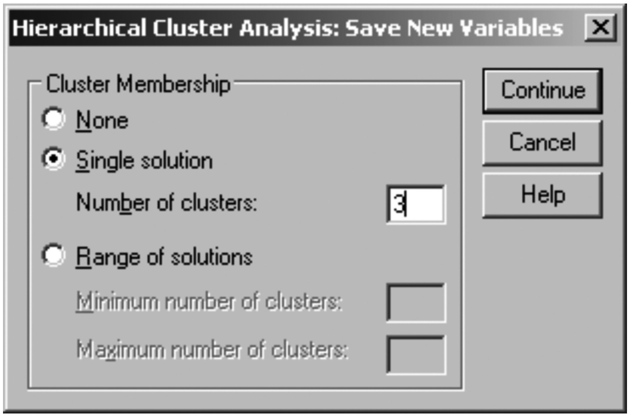


Figure 23.4 **Dendrogram using Ward method.**



Hierarchical Cluster Analysis: Save New Variables

Cluster Membership

○ None

● Single solution

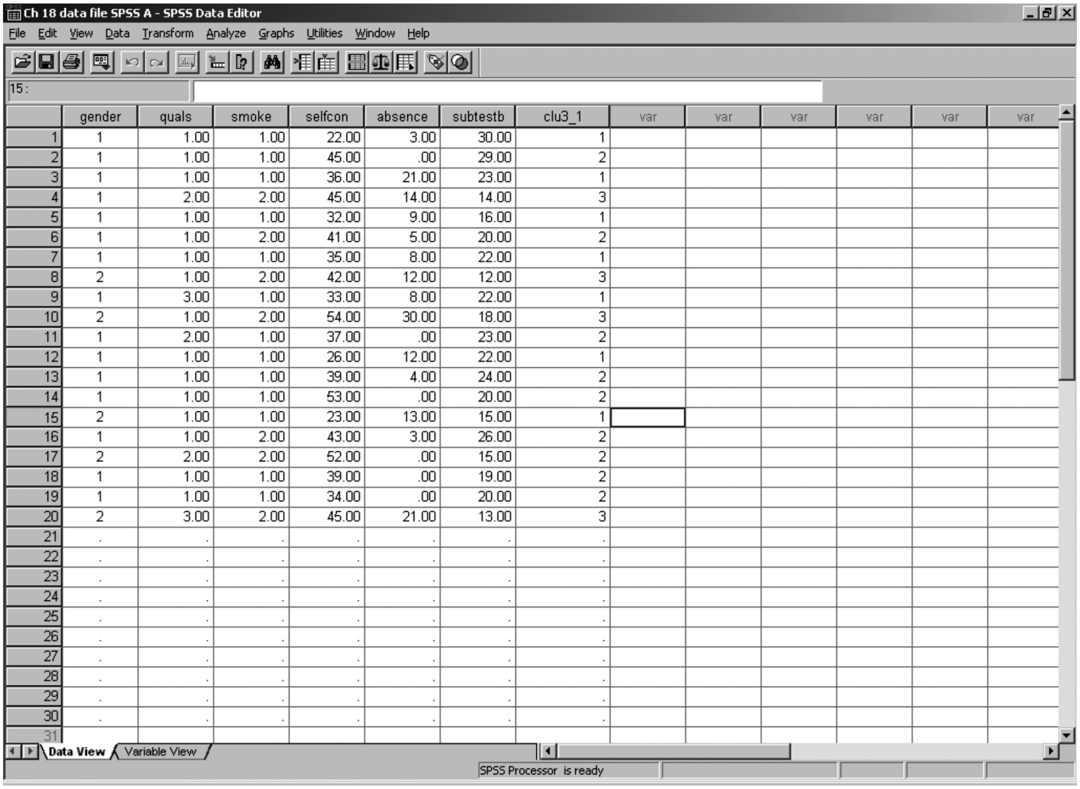
Number of clusters: 3

○ Range of solutions

Minimum number of clusters:

Maximum number of clusters:

Figure 23.5 **Save new variables box.**


Ch 18 data file SPSS A - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
15 :
Data View
Variable View
SPSS Processor is ready

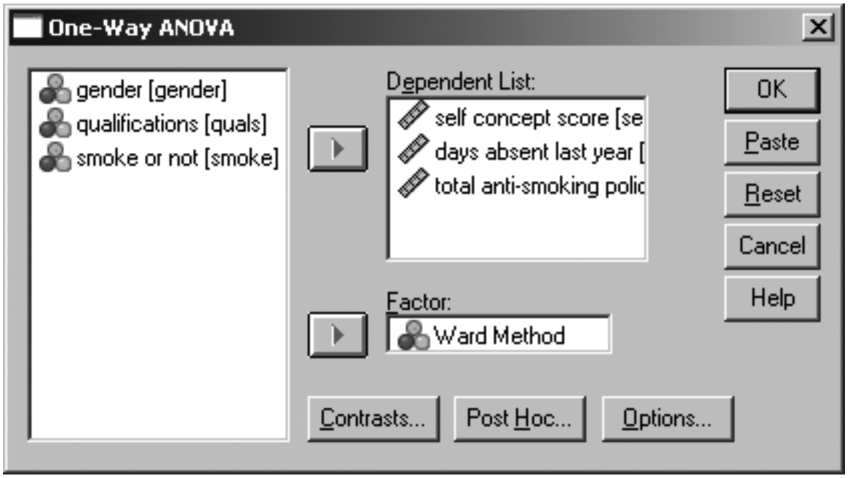
	gender	quals	smoke	selfcon	absence	subtestb	clu3_1	var	var	var	var	var	var
1	1	1.00	1.00	22.00	3.00	30.00	1						
2	1	1.00	1.00	45.00	.00	29.00	2						
3	1	1.00	1.00	36.00	21.00	23.00	1						
4	1	2.00	2.00	45.00	14.00	14.00	3						
5	1	1.00	1.00	32.00	9.00	16.00	1						
6	1	1.00	2.00	41.00	5.00	20.00	2						
7	1	1.00	1.00	35.00	8.00	22.00	1						
8	2	1.00	2.00	42.00	12.00	12.00	3						
9	1	3.00	1.00	33.00	8.00	22.00	1						
10	2	1.00	2.00	54.00	30.00	18.00	3						
11	1	2.00	1.00	37.00	.00	23.00	2						
12	1	1.00	1.00	26.00	12.00	22.00	1						
13	1	1.00	1.00	39.00	4.00	24.00	2						
14	1	1.00	1.00	53.00	.00	20.00	2						
15	2	1.00	1.00	23.00	13.00	15.00	1						
16	1	1.00	2.00	43.00	3.00	26.00	2						
17	2	2.00	2.00	52.00	.00	15.00	2						
18	1	1.00	1.00	39.00	.00	19.00	2						
19	1	1.00	1.00	34.00	.00	20.00	2						
20	2	3.00	2.00	45.00	21.00	13.00	3						
21	.	.	.	.	.	.	.						
22	.	.	.	.	.	.	.						
23	.	.	.	.	.	.	.						
24	.	.	.	.	.	.	.						
25	.	.	.	.	.	.	.						
26	.	.	.	.	.	.	.						
27	.	.	.	.	.	.	.						
28	.	.	.	.	.	.	.						
29	.	.	.	.	.	.	.						
30	.	.	.	.	.	.	.						
31													

Figure 23.6 **New variable clu_1.**

Nine respondents have been classified in cluster 2, while there are seven in cluster 1 and four in cluster 3. Normally, we now proceed by conducting a one-way ANOVA to determine on which classifying variables are significantly different between the groups because there would be a large number of respondents. With only 20 cases in this example, an ANOVA is not really feasible. However, for the purposes of demonstration, we will do one to show how it helps with determining what each cluster is based on.

In conducting the one-way ANOVA, you would calculate the descriptives on the scale (interval) variables for each of the clusters and note the differences. The grouping variable is the new clusters variable. Categorical (nominal) data can be dealt with using Crosstabs (see below). The One-Way ANOVA box (Fig. 23.7) and descriptives table (Table 23.3) are shown below.

There appear to be some major differences between the means of various clusters for each variable in Table 23.3. These are explored in ANOVA Table 23.4. which offers F values and significance levels to show whether any of these mean differences are significant. The between groups means are all significant, indicating each of the three variables reliably distinguish between the three clusters. With a significant ANOVA and three or more clusters, as in this example, a Tukey *post-hoc* test is also necessary to determine where the differences lie.



One-Way ANOVA

gender [gender]
qualifications [quals]
smoke or not [smoke]

Dependent List:
self concept score [se
days absent last year [
total anti-smoking polic

Factor:
Ward Method

Figure 23.7 **One-way ANOVA box.**

Table 23.3 **Descriptives table**

		N	Mean	Std. deviation	Std. error	Minimum	Maximum
self concept score	1	7	29.5714	5.79819	2.19151	22.00	36.00
	2	9	42.5556	6.48288	2.16096	34.00	53.00
	3	4	46.5000	5.19615	2.59808	42.00	54.00
	Total	20	38.8000	9.11679	2.03858	22.00	54.00
days absent last year	1	7	10.5714	5.62308	2.12533	3.00	21.00
	2	9	1.3333	2.06155	.68718	.00	5.00
	3	4	19.2500	8.13941	4.06971	12.00	30.00
	Total	20	8.1500	8.50557	1.90190	.00	30.00
total anti-smoking policies subtest B	1	7	21.4286	4.96176	1.87537	15.00	30.00
	2	9	21.7778	4.17665	1.39222	15.00	29.00
	3	4	14.2500	2.62996	1.31498	12.00	18.00
	Total	20	20.1500	5.03958	1.12688	12.00	30.00

The Tukey post *hoc-test* (Table 23.5) reveals that self-concept score and days absent reliably differentiate three clusters through their cluster means. Anti-smoking policy attitudes only significantly differentiate between clusters 2 and 3 and 1 and 3. Clusters 1 and 2 are not significantly different on this variable.

Crosstab analysis of the nominal variables *gender*, *qualifications* and *whether smoke or not* produced some significant associations with clusters. Of course, with such small numbers we would not normally have conducted a crosstabs, as many cells would have counts less than 5. This explanation has been solely to show how you can tease out the characteristics of the clusters.

Table 23.4 **ANOVA table**

		Sum of squares	df	Mean square	F	Sig.
self concept score	Between Groups	960.263	2	480.132	13.188	.000
	Within Groups	618.937	17	36.408		
	Total	1579.200	19			
days absent last year	Between Groups	952.086	2	476.043	19.156	.000
	Within Groups	422.464	17	24.851		
	Total	1374.550	19			
total anti-smoking policies subtest B	Between Groups	174.530	2	87.265	4.816	.022
	Within Groups	308.020	17	18.119		
	Total	482.550	19			

Table 23.5 **Multiple comparisons**

Tukey HSD

						95% Confidence interval	
Dependent variable	(I) Ward method	(J) Ward method	Mean difference (I-J)	Std. error	Sig.	Lower bound	Upper bound
self concept score	1	2	–12.98413(*)	3.04080	.001	–20.7849	–5.1834
		3	–16.92857(*)	3.78195	.001	–26.6306	–7.2265
	2	1	12.98413(*)	3.04080	.001	5.1834	20.7849
		3	–3.94444	3.62593	.534	–13.2462	5.3574
	3	1	16.92857(*)	3.78195	.001	7.2265	26.6306
		2	3.94444	3.62593	.534	–5.3574	13.2462
days absent last year	1	2	9.23810(*)	2.51223	.005	2.7933	15.6829
		3	–8.67857(*)	3.12455	.033	–16.6942	–.6630
	2	1	–9.23810(*)	2.51223	.005	–15.6829	–2.7933
		3	–17.91667(*)	2.99565	.000	–25.6016	–10.2318
	3	1	8.67857(*)	3.12455	.033	.6630	16.6942
		2	17.91667(*)	2.99565	.000	10.2318	25.6016
total anti-smoking policies subtest B	1	2	–.34921	2.14513	.986	–5.8522	5.1538
		3	7.17857(*)	2.66798	.039	.3343	14.0229
	2	1	.34921	2.14513	.986	–5.1538	5.8522
		3	7.52778(*)	2.55791	.023	.9658	14.0897
	3	1	–7.17857(*)	2.66798	.039	–14.0229	–.3343
		2	–7.52778(*)	2.55791	.023	–14.0897	–.9658

* The mean difference is significant at the .05 level.

The three clusters significantly differentiated between *gender* with males in 1 and 2 and females in 3. *Smoking* also produced significant associations with *non-smokers* in 1 and *smokers* in 3. Cluster 2 represented both *smokers* and *non-smokers* who were differentiated by other variables. *Qualifications* did not produce any significant associations. Histograms can be produced with the crosstabs analysis to reveal useful visual differentiations of the groupings.

When cluster memberships are significantly different they can be used as a new grouping variable in other analyses. The significant differences between variables for the clusters suggest the ways in which the clusters differ or on which they are based. In our example:

- Cluster 1 is characterized by low self-concept, average absence rate, average attitude score to anti-smoking, non-smoking males.
- Cluster 2 is characterized by moderate self-concept, low absence rate, average attitude score to anti-smoking, smoking and non-smoking males.
- Cluster 3 is characterized by high self-concept, high absence rate, low score to anti-smoking, smoking females.

It is important to remember that cluster analysis will always produce a grouping, but these may or may not prove useful for classifying items.

Write up of results

'A cluster analysis was run on 20 cases, each responding to items on demographics (gender; qualifications, days absence from work, whether they smoke or not), on their attitudes to smoking in public places, and score for self-concept scale. A hierarchical cluster analysis using Ward's method produced three clusters, between which the variables were significantly different in the main. The first cluster was predominant and characterized by non-smoking low self-concept males. The third cluster was essentially high self-concept pro-smoking females. The middle cluster was again mainly male, both smoking and non-smoking, with average positions on the other variables'.

SPSS Activity

Now access Ch. 23 data file SPSS B on the website and conduct your own cluster analysis. This file contains mean scores on five attitude scales, age and salary data for 20 respondents. Write out an explanation of the results and discuss in class.

What you have learned from this chapter

Cluster analysis determines how many 'natural' groups there are in the sample. It also allows you to determine who in your sample belongs to which group. Cluster analysis is not as much a typical statistical test as it is a 'collection' of different algorithms that put objects into clusters according to well-defined similarity rules. The aim is to (1) minimize variability within clusters and (2) maximize variability between clusters. The most common approach is to use hierarchical cluster analysis and Ward's method.

Unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any 'a priori' hypotheses, but are still in the exploratory phase of research. Once clear clusters are identified the analysis is re-rerun with only those clusters to allocate items/respondents to the selected clusters. This forms a new variable on the SPSS data view that enables existing and new items/respondents to be classified. Significant differences between clusters can be tested with ANOVA or a non-parametric test.

Review questions

Qu. 23.1
Cluster analysis attempts to:

(a) determine the best predictors for an interval DV
(b) determine the optimum groups differentiated by the IV's
(c) determine the best predictors for categorical DV's
(d) determine the minimum number of groups that explain the DV
(e) none of the above

Qu. 23.2
Why do you run a One-Way ANOVA after assigning cases to clusters in cluster analysis?

(a) to determine if there are any significant cases
(b) to determine the defining characteristics of each cluster
(c) to determine if p is significant
(d) to determine which variables significantly differentiate between clusters
(e) all of the above

Qu. 23.3
Explain the difference between Ward's method and *k* means clustering.

Check your response with the material above.

Now access the Web page for Chapter 23 and check your answers to the above questions. You should also attempt the SPSS activity.