Hierarchical Cluster Analysis

Contents

Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. It is most useful when you want to cluster a small number (less than a few hundred) of objects.

The **objects** in hierarchical cluster analysis can be cases or variables, depending on whether you want to classify cases or examine relationships between the variables.

Next

Clustering Principles

Using HCA to Classify Automobiles

Using HCA to Study Relationships between Telecommunications Services

Related Procedures

Recommended Readings

1. Clustering Principles

Hierarchical cluster analysis begins by separating each object into a cluster by itself. At each stage of the analysis, the criterion by which objects are separated is relaxed in order to link the two most similar clusters until all of the objects are joined in a complete classification tree.

The basic criterion for any clustering is distance. Objects that are near each other should belong to the same cluster, and objects that are far from each other should belong to different clusters. For a given set of data, the clusters that are constructed depend on your specification of the following parameters:

- **Cluster method** defines the rules for cluster formation. For example, when calculating the distance between two clusters, you can use the pair of nearest objects between clusters or the pair of furthest objects, or a compromise between these methods.
- **Measure** defines the formula for calculating distance. For example, the Euclidean distance measure calculates the distance as a "straight line" between two clusters. Interval measures assume that the variables are scale; count measures assume that they are discrete numeric; and binary measures assume that they take only two values.
- **Standardization** allows you to equalize the effect of variables measured on different scales.

Next

Parent topic: Hierarchical Cluster Analysis

2. Using HCA to Classify Automobiles

Car manufacturers need to be able to appraise the current market to determine the likely competition for their vehicles. If cars can be grouped according to available data, this task can be largely automatic using cluster analysis.

Information for various makes and models of motor vehicles is contained in *car_sales.sav*. See the topic Sample Files for more information. Use the Hierarchical Cluster Analysis procedure to group the highest-selling automobiles according to their prices and physical properties.

Next

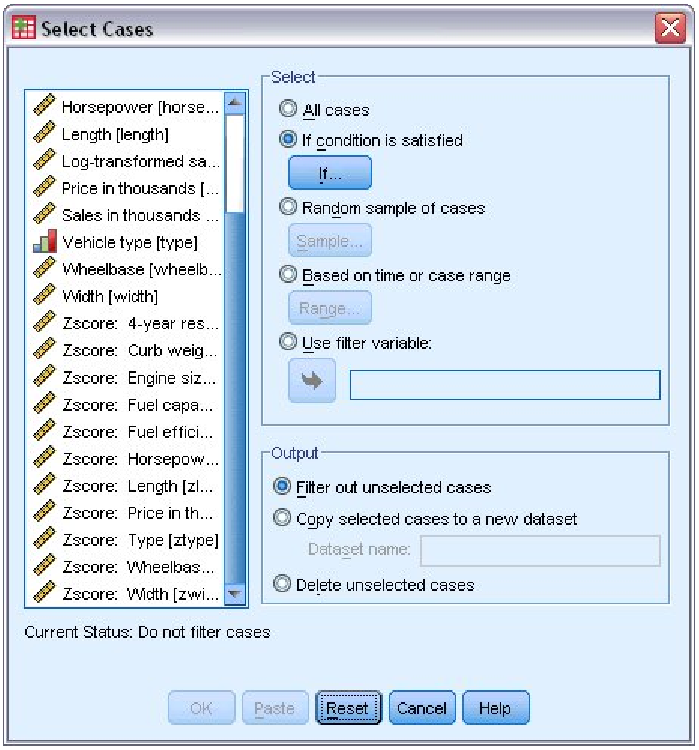
Parent topic: Hierarchical Cluster Analysis

2.1. Preparing the Data

1. To select the cases for analysis, from the menus choose:

Data > **Select Cases...**

Figure 1. Select Cases main dialog box



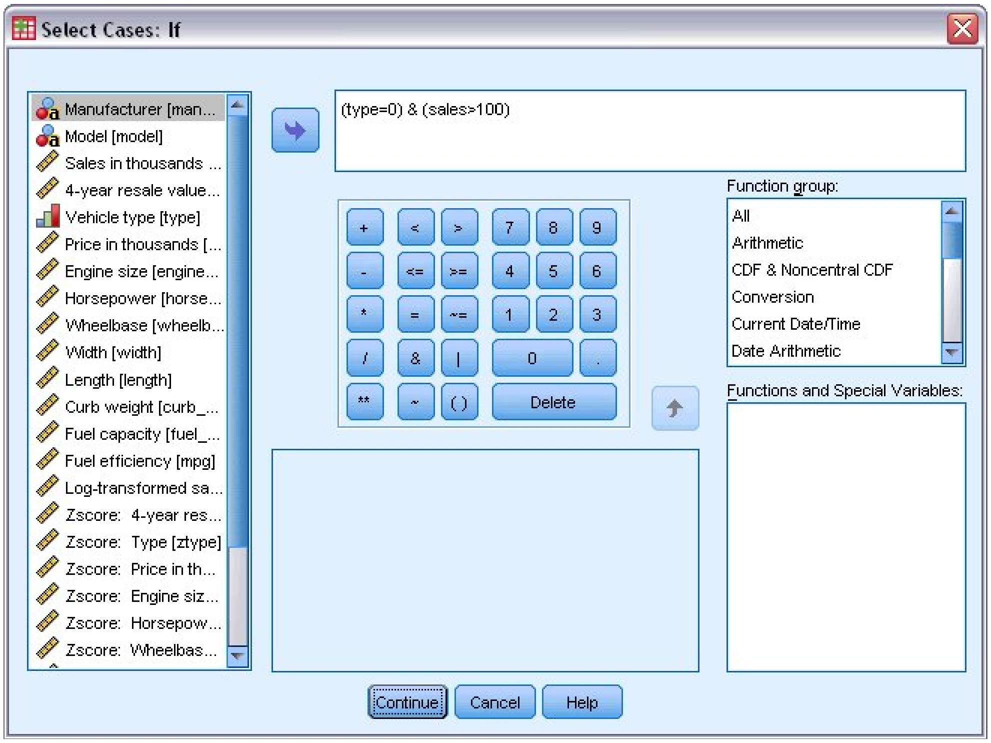
Select
○ All cases
● If condition is satisfied
○ Random sample of cases
○ Based on time or case range
○ Use filter variable:

Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

2. Select **If condition is satisfied**.
3. Click **If**.

Figure 2. Select Cases If dialog box



(type=0) & (sales>100)

4. In the text field, type `(type=0) & (sales>100)`.
5. Click **Continue**.
6. Click **OK** in the Select Cases dialog box.

Next

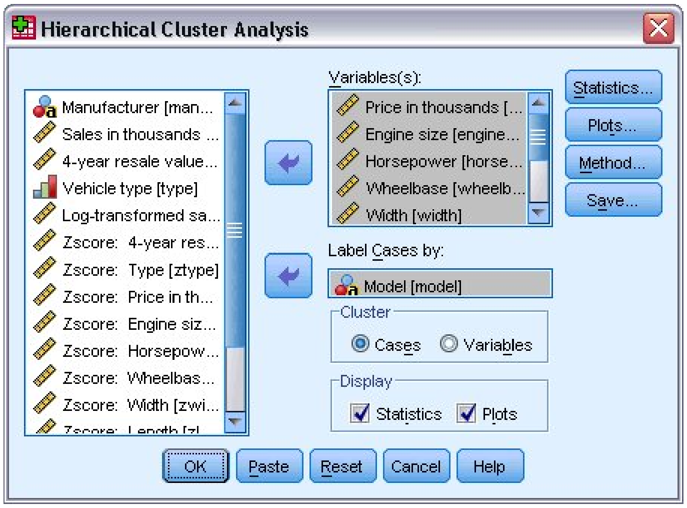
Parent topic: Using HCA to Classify Automobiles

2.2. Running the Analysis

1. To run the cluster analysis, from the menus choose:

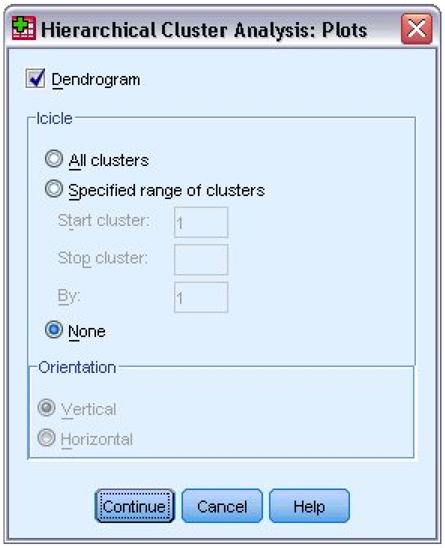
   Analyze > **Classify** > **Hierarchical Cluster...**

   Figure 1. Hierarchical Cluster Analysis dialog box



2. If the variable list does not display variable labels in file order, right-click anywhere in the variable list and from the context menu choose **Display Variable Labels** and **Sort by File Order**.
3. Select *Price in thousands* through *Fuel efficiency* as analysis variables.
4. Select *Model* as the case labeling variable.
5. Click **Plots**.

   Figure 2. Plots dialog box



Hierarchical Cluster Analysis: Plots

☑ Dendrogram

Icicle

○ All clusters

○ Specified range of clusters

Start cluster: 1

Stop cluster:

By: 1

● None

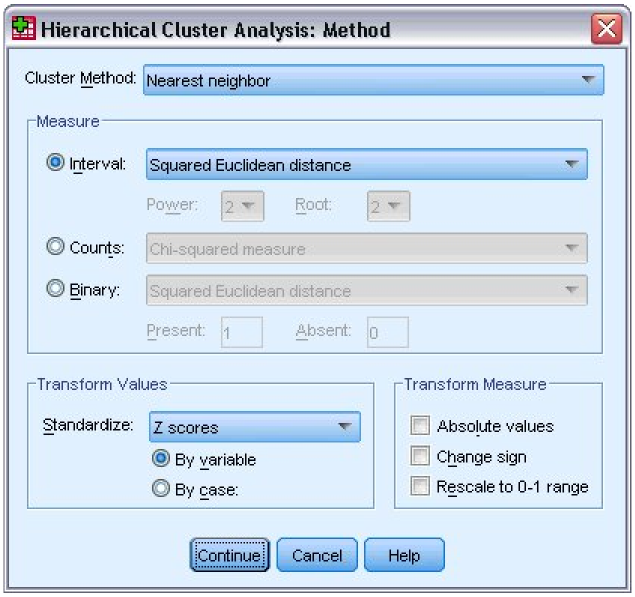
Orientation

● Vertical

○ Horizontal

6. Select **Dendrogram**.
7. Select **None** in the Icicle group.
8. Click **Continue**.
9. Click **Method** in the Hierarchical Cluster Analysis dialog box.

Figure 3. Method dialog box



Hierarchical Cluster Analysis: Method

Cluster Method: Nearest neighbor

Measure

● Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: Z scores

● By variable

○ By case:

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

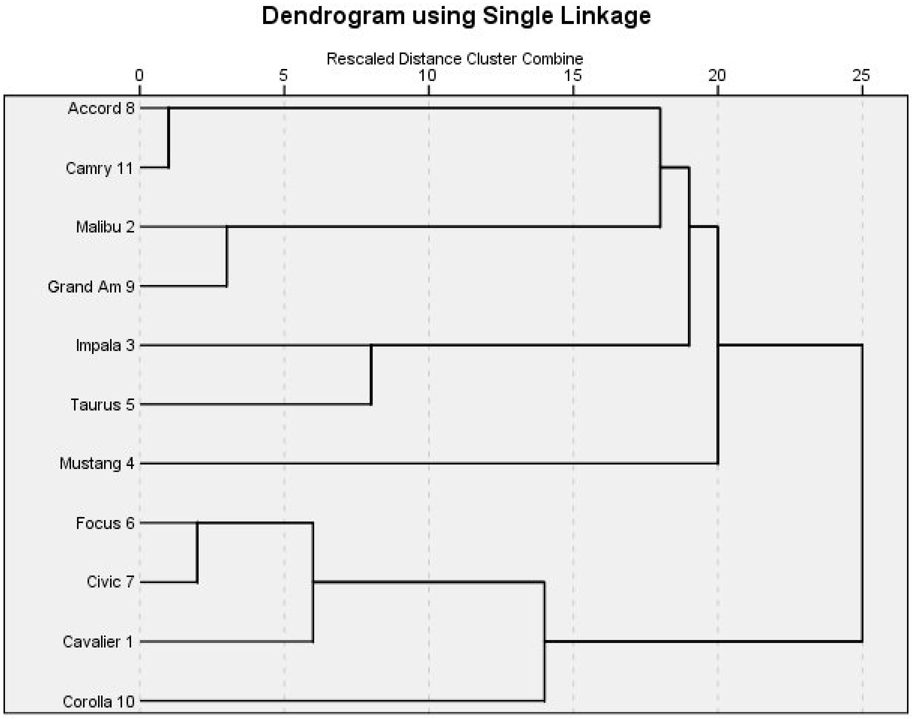
10. Select **Nearest neighbor** as the cluster method.
11. Select **Z scores** as the standardization in the Transform Values group.
12. Click **Continue**.
13. Click **OK** in the Hierarchical Cluster Analysis group.

Next

Parent topic: Using HCA to Classify Automobiles

2.3. Dendrogram

Figure 1. Dendrogram for single linkage solution



The dendrogram is a graphical summary of the cluster solution. Cases are listed along the left vertical axis. The horizontal axis shows the distance between clusters when they are joined.

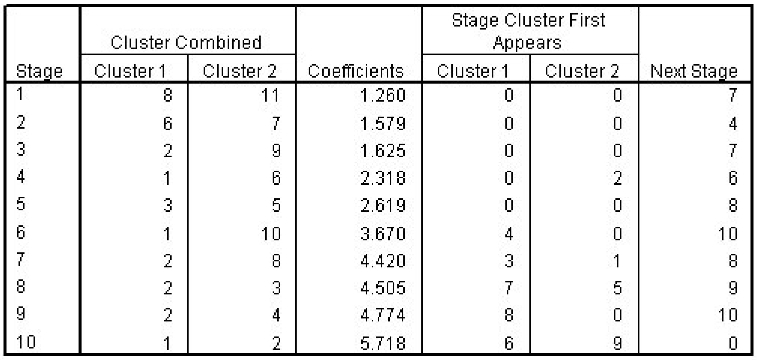
Parsing the classification tree to determine the number of clusters is a subjective process. Generally, you begin by looking for "gaps" between joinings along the horizontal axis. Starting from the right, there is a gap between 20 and 25, which splits the automobiles into two clusters. There is another gap from approximately 10 to 15, which suggests 6 clusters.

Next

Parent topic: Using HCA to Classify Automobiles

2.4. Agglomeration Schedule

Figure 1. Agglomeration schedule for single linkage solution



Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

The agglomeration schedule is a numerical summary of the cluster solution. At the first stage, cases 8 and 11 are combined because they have the smallest distance. The cluster created by their joining next appears in stage 7. In stage 7, the clusters created in stages 1 and 3 are joined. The resulting cluster next appears in stage 8. When there are many cases, this table becomes rather long, but it may be easier to scan the coefficients column for large gaps rather than scan the dendrogram.

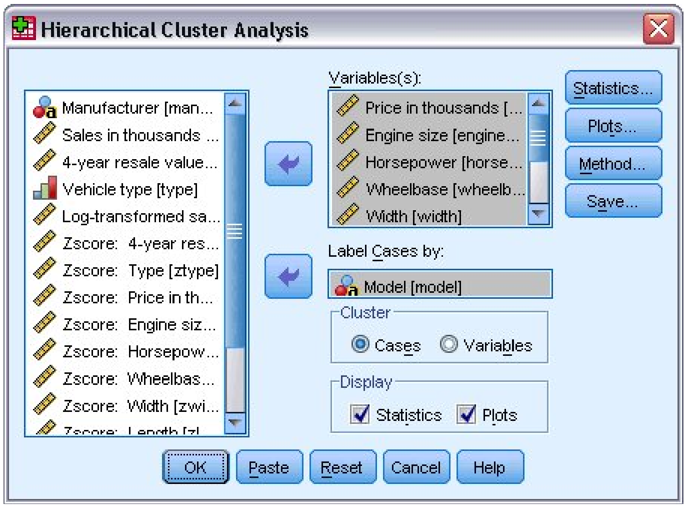
A good cluster solution sees a sudden jump (gap) in the distance coefficient. The solution before the gap indicates the good solution. The largest gaps in the coefficients column occur between stages 5 and 6, indicating a 6-cluster solution, and stages 9 and 10, indicating a 2-cluster solution. These are the same as your findings from the dendrogram. This is somewhat unsatisfactory as a solution, because there isn't a strong classification. Try a solution using complete linkage (Furthest neighbor) as the cluster method.

Next

Parent topic: Using HCA to Classify Automobiles

2.5. Complete Linkage Solution

Figure 1. Hierarchical Cluster Analysis



Variables(s):
- Price in thousands [...
- Engine size [engine...
- Horsepower [horse...
- Wheelbase [wheelb...
- Width [width]

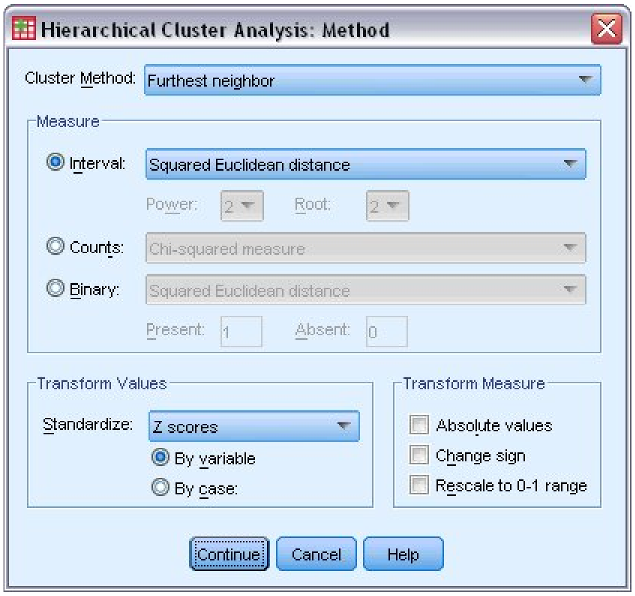
Label Cases by: Model [model]

Cluster: ● Cases ○ Variables

Display: ☑ Statistics ☑ Plots

1. To run a cluster analysis using complete linkage, recall the Hierarchical Cluster dialog box.
2. Click **Method**.

   Figure 2. Method dialog box



Hierarchical Cluster Analysis: Method

Cluster Method: Furthest neighbor

Measure

● Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: Z scores

● By variable

○ By case:

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

3. Select **Furthest neighbor** as the cluster method.
4. Click **Continue**.
5. Click **OK** in the Hierarchical Cluster Analysis dialog box.

Next

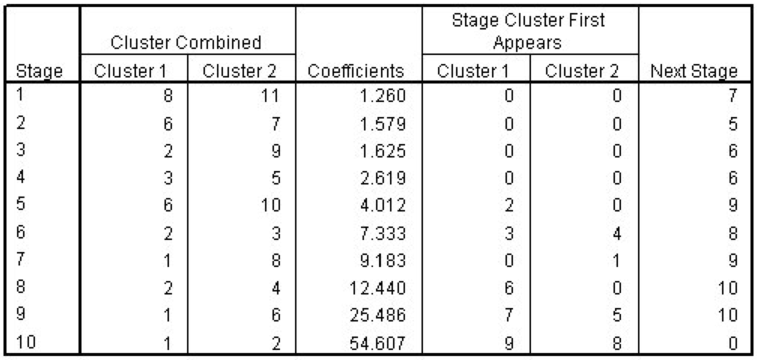
Agglomeration Schedule

Dendrogram

Parent topic: Using HCA to Classify Automobiles

2.5.1. Agglomeration Schedule

Figure 1. Agglomeration schedule for complete linkage solution



Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	5
3	2	9	1.625	0	0	6
4	3	5	2.619	0	0	6
5	6	10	4.012	2	0	9
6	2	3	7.333	3	4	8
7	1	8	9.183	0	1	9
8	2	4	12.440	6	0	10
9	1	6	25.486	7	5	10
10	1	2	54.607	9	8	0

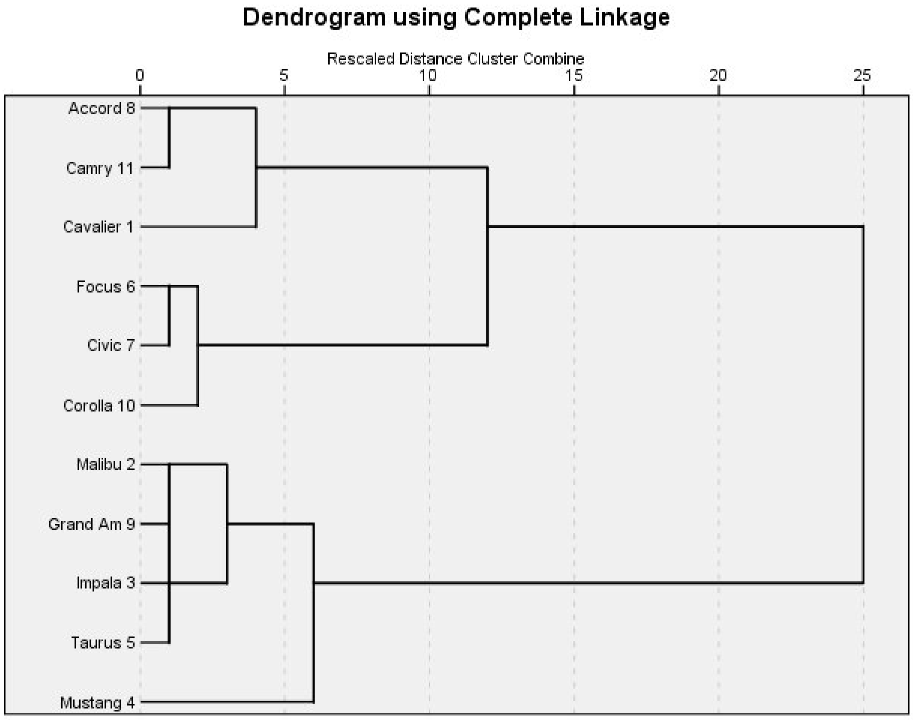
For the first few stages, the schedule for the complete linkage solution is similar to that for the single linkage solution. In the final few stages, they are quite different as the complete linkage solution makes a strong classification of two or three clusters.

Next

Parent topic: Complete Linkage Solution

2.5.2. Dendrogram

Figure 1. Dendrogram for complete linkage solution



The decisiveness of this classification is reflected in the dendrogram. The initial splitting of the tree forms two clusters. The top contains smaller cars. The bottom contains larger cars. The cluster of smaller cars can be further split into small and economy cars. The Civic and Corolla are both smaller and cheaper siblings of the Accord and Camry, respectively.

Next

Parent topic: Complete Linkage Solution

2.6. Summary

The complete linkage solution is satisfying because its clusters are distinct, while the single linkage solution is less conclusive. Using complete linkage clustering, you can determine the competition for vehicles in the design phase by entering their specifications as new cases in the data set and rerunning the analysis.

Next

Parent topic: Using HCA to Classify Automobiles

3. Using HCA to Study Relationships between Telecommunications Services

A telecommunications provider wants to better understand service usage patterns in its customer base. If services can be clustered by usage, the company can offer more attractive packages to its customers.

Variables indicating usage and non-usage of services are contained in *telco.sav*. See the topic Sample Files for more information. Use the Hierarchical Cluster Analysis procedure to study the relationships between the various services.

Next

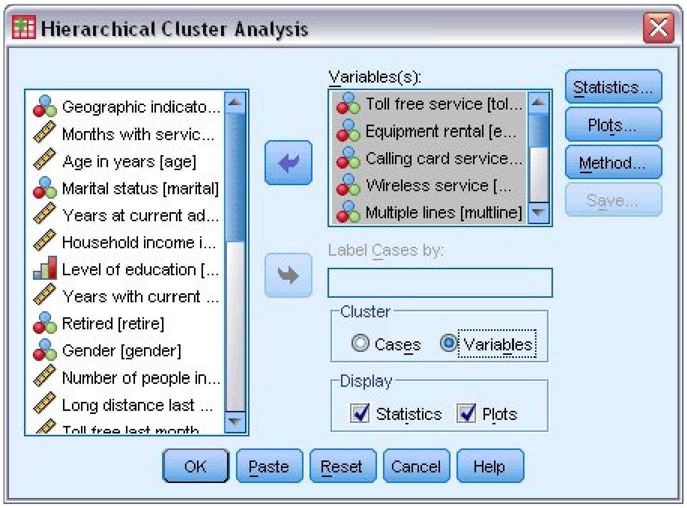
Parent topic: Hierarchical Cluster Analysis

3.1. Running the Analysis

1. To run the cluster analysis, from the menus choose:

   Analyze > **Classify** > **Hierarchical Cluster...**

   Figure 1. Hierarchical Cluster Analysis dialog box



Hierarchical Cluster Analysis

Geographic indicato...
Months with servic...
Age in years [age]
Marital status [marital]
Years at current ad...
Household income i...
Level of education [...
Years with current ...
Retired [retire]
Gender [gender]
Number of people in...
Long distance last ...
Toll free last month

Variables(s):
Toll free service [tol...
Equipment rental [e...
Calling card service...
Wireless service [...
Multiple lines [multline]

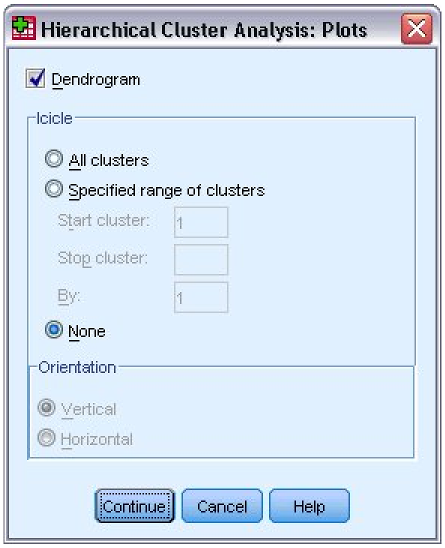
Label Cases by:

Cluster: ○ Cases ● Variables

Display: ☑ Statistics ☑ Plots

2. Click **Reset** to restore the default settings.
3. If the variable list does not display variable labels in file order, right-click anywhere in the variable list and from the context menu choose **Display Variable Labels** and **Sort by File Order**.
4. Select *Toll free service* through *Wireless service* and *Multiple lines* through *Electronic billing* as analysis variables.
5. Select **Variables** in the Cluster group.
6. Click **Plots**.

   Figure 2. Plots dialog box



Hierarchical Cluster Analysis: Plots

☑ Dendrogram

Icicle

○ All clusters

○ Specified range of clusters

Start cluster: 1

Stop cluster:

By: 1

● None

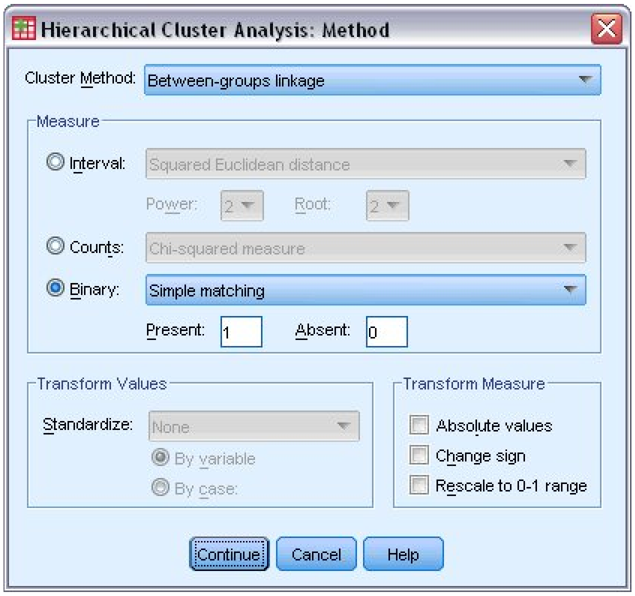
Orientation

● Vertical

○ Horizontal

7. Select **Dendrogram**.
8. Select **None** in the Icicle group.
9. Click **Continue**.
10. Click **Method** in the Hierarchical Cluster Analysis dialog box.

Figure 3. Method dialog box



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

○ Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

● Binary: Simple matching

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case:

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

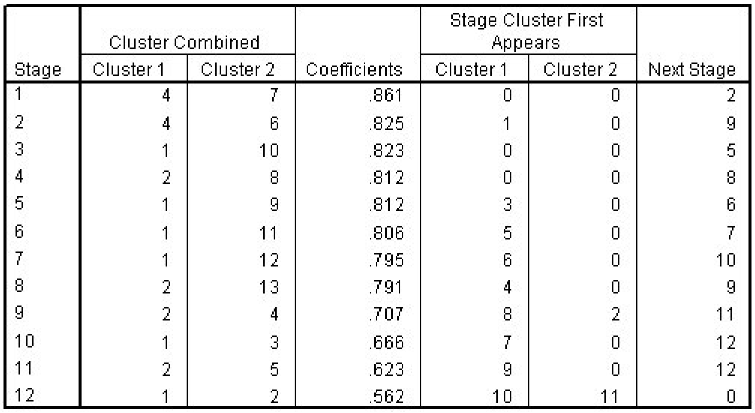
11. Select **Binary** in the Measure group.
12. Select **Simple Matching** as the binary measure.
13. Click **Continue**.
14. Click **OK** in the Hierarchical Cluster Analysis dialog box.

Next

Parent topic: Using HCA to Study Relationships between Telecommunications Services

3.2. Agglomeration Schedule

Figure 1. Agglomeration schedule for simple matching solution



Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	7	.861	0	0	2
2	4	6	.825	1	0	9
3	1	10	.823	0	0	5
4	2	8	.812	0	0	8
5	1	9	.812	3	0	6
6	1	11	.806	5	0	7
7	1	12	.795	6	0	10
8	2	13	.791	4	0	9
9	2	4	.707	8	2	11
10	1	3	.666	7	0	12
11	2	5	.623	9	0	12
12	1	2	.562	10	11	0

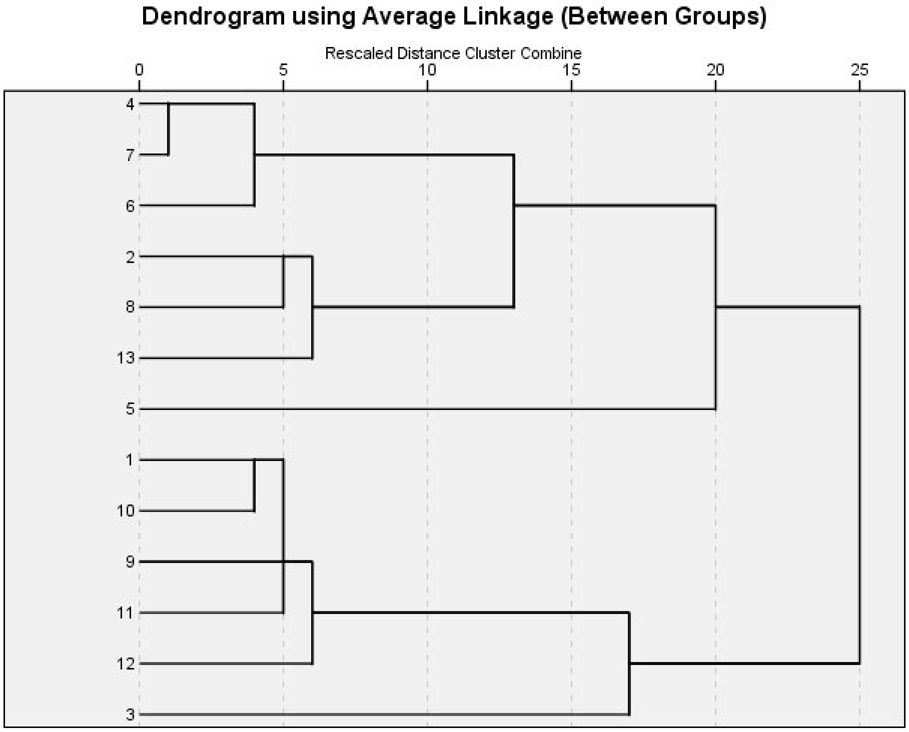
For binary measures, the coefficients column reports measures of similarity, thus the coefficient values decrease at each stage of the analysis. It is difficult to interpret the results of this schedule, so turn to the dendrogram.

Next

Parent topic: Using HCA to Study Relationships between Telecommunications Services

3.3. Dendrogram

Figure 1. Dendrogram for simple matching solution



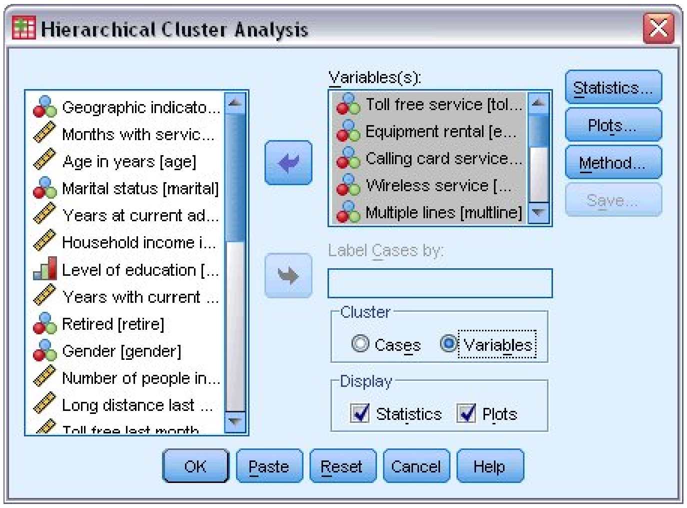
The dendrogram shows that the usage patterns of *MULTLINE* and *CALLCARD* are different from the other services. These others are grouped into three clusters. One cluster includes *WIRELESS*, *PAGER*, and *VOICE*. Another includes *EQUIP*, *INTERNET*, and *EBILL*. The last contains *TOLLFREE*, *CALLWAIT*, *CALLID*, *FORWARD*, and *CONFER*. The *WIRELESS* cluster is closer to the *INTERNET* cluster than the *CALLWAIT* cluster.

Next

Parent topic: Using HCA to Study Relationships between Telecommunications Services

3.4. Solution Using the Jaccard Measure

Figure 1. Hierarchical Cluster Analysis dialog box



Variables(s):
Toll free service [tol...
Equipment rental [e...
Calling card service...
Wireless service [...
Multiple lines [multline]

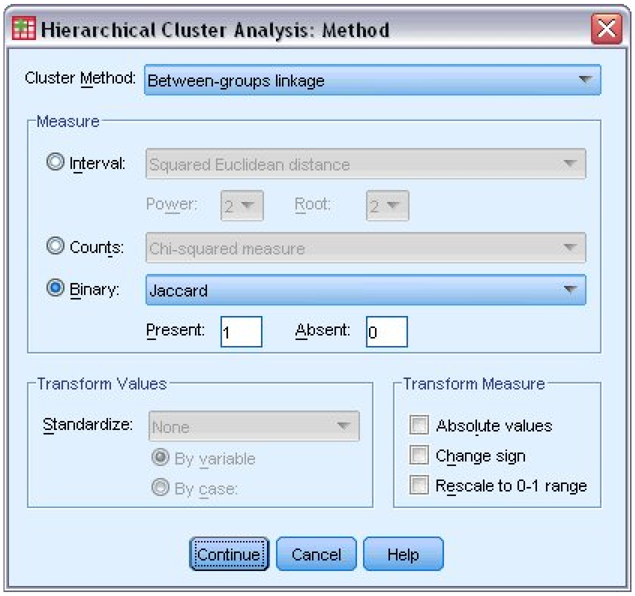
Label Cases by:

Cluster: ○ Cases ● Variables

Display: ☑ Statistics ☑ Plots

1. To run a cluster analysis using the Jaccard distance measure, recall the Hierarchical Cluster dialog box.
2. Click **Method**.

   Figure 2. Method dialog box



Cluster Method: Between-groups linkage

Measure

○ Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

● Binary: Jaccard

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

3. Select **Jaccard** as the binary measure.
4. Click **Continue**.
5. Click **OK** in the Hierarchical Cluster Analysis dialog box.

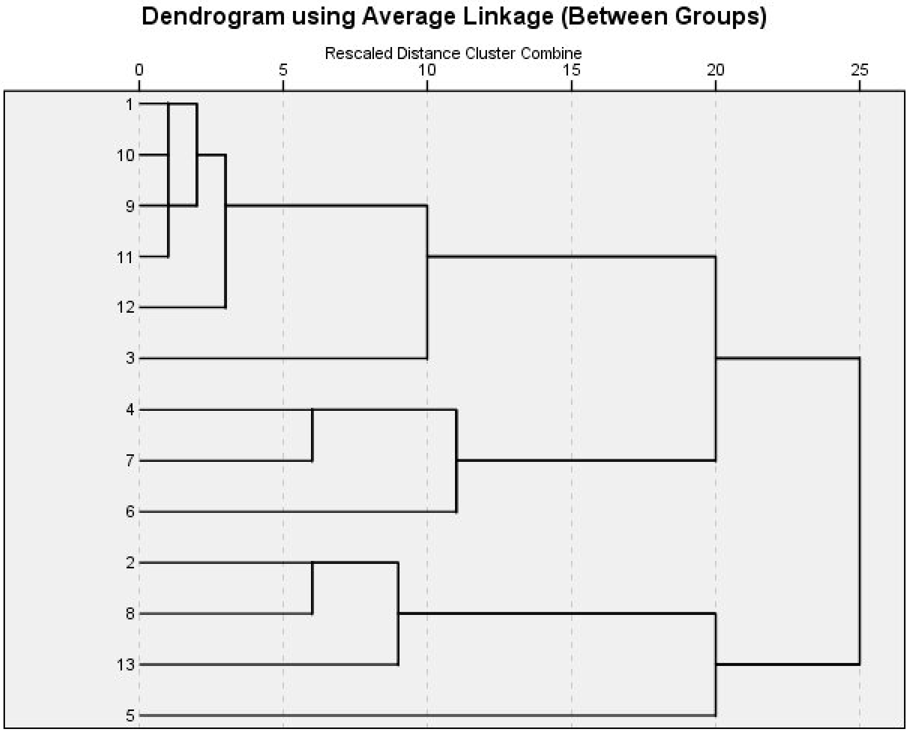
Next

Dendrogram

Parent topic: Using HCA to Study Relationships between Telecommunications Services

3.4.1. Dendrogram

Figure 1. Dendrogram for Jaccard solution



Using the Jaccard measure, the three basic clusters are the same, but the *WIRELESS* cluster is now closer to the *CALLWAIT* cluster than the *INTERNET* cluster.

Next

Parent topic: Solution Using the Jaccard Measure

3.5. Summary

The difference between the simple matching and Jaccard measures is that the Jaccard measure does not consider two services to which an individual does not subscribe to be similar. Therefore, simple matching considers wireless and Internet services to be similar when a customer either has both or neither, while Jaccard considers them to be similar only when a customer has both. This causes a difference in the cluster solutions because there are many customers who don't have wireless or Internet services. Thus, these clusters are more similar in the simple matching solution than the Jaccard solution. The measure that you use depends on the definition of "similar" that applies to your situation.

Next

Parent topic: Using HCA to Study Relationships between Telecommunications Services

4. Related Procedures

The Hierarchical Cluster Analysis procedure is useful for finding natural groupings of cases or variables. It works best when your data file contains a small number (less than a few hundred) of objects to be clustered.

- If you have a large number of cases to be clustered, use the TwoStep Cluster Analysis procedure.
- If you have a large number of cases to be clustered and all your variables are scale, you can alternately use the K-Means Cluster Analysis procedure.
- If you're interested in examining the structure of your variables and they are scale variables, you can try Factor Analysis as an alternative.

Next

Parent topic: Hierarchical Cluster Analysis

5. Recommended Readings

See the following texts for more information on hierarchical cluster analysis:

Aldenderfer, M. S., and R. K. Blashfield. 1984. *Cluster Analysis*. Newbury Park: Sage Publications.

Parent topic: Hierarchical Cluster Analysis