

# Chapter 8 Cluster Analysis

## SPSS - Cluster Analysis

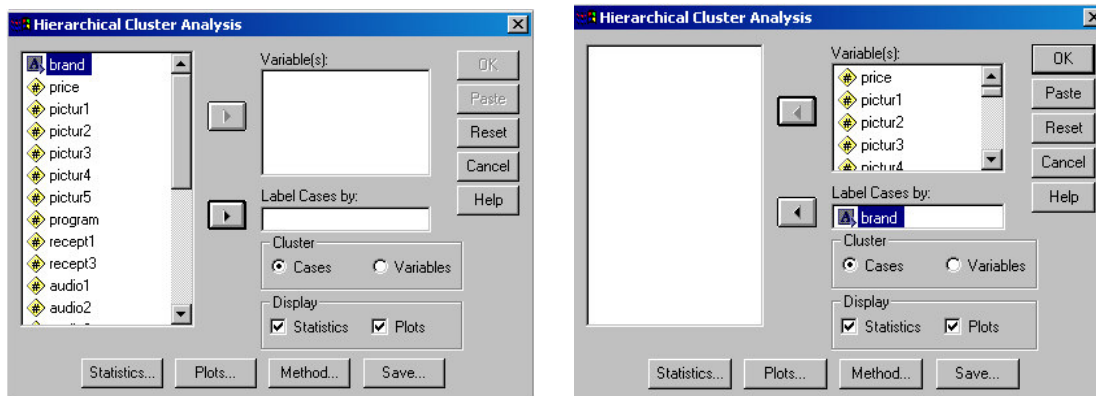
Datafile used: *vcr.sav*. This datafile is about the quality of the 21, fictional, brands of VCRs.

**How to get there: Analyze → Classify → ...**

### **→ Hierarchical Cluster...**

This procedure attempts to identify relatively homogeneous groups of cases (or variables) based on selected characteristics. For example: cluster television shows into homogeneous groups based on viewer characteristics. In hierarchical clustering, an algorithm is used that starts with each case (or variable) in a separate cluster and combines clusters until only one is left.

This menu selection opens the following Hierarchical Clustering Analysis main dialog window.

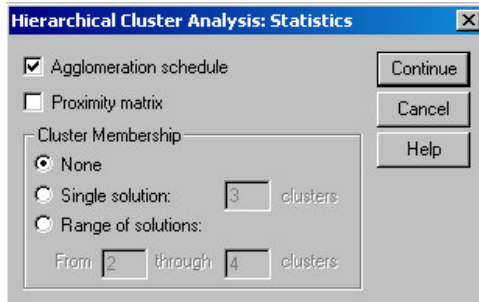


The use of this window will vary substantially depending on whether you choose to Cluster Cases or to Cluster Variables; here we focus on **Clustering Cases**.

To cluster cases you need to identify variables you wish to be considered in creating clusters for the cases. The variables to be used for cluster formation are here: picture quality (5 measures), reception quality (3 measures), audio quality (3 measures), ease of programming (1 measure), number of events (1 measure), number of days for future programming (1 measure), remote control (3 measures), and extras (3 measures). Pass these in the Variable(s) box.

Then you need to specify how you wish your cases to be identified. This will usually be an ID number or some identifying name (here the variable *brand*). See the second picture above.

### **Button → Statistics ...**



Here you can define the Cluster membership:

**None:** All clusters are listed since all possible solutions will be identified. What it does *not* do is identify cases included in each cluster for a particular solution.

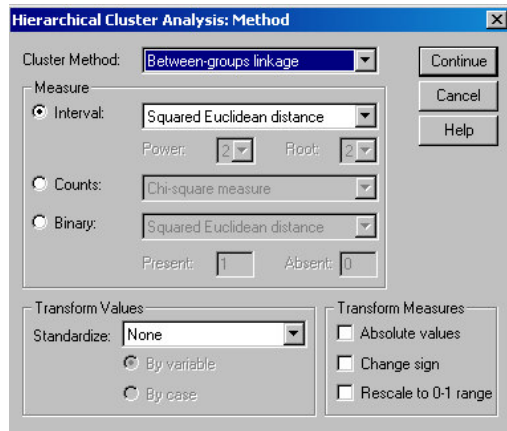
**Single solution:** Specify some number greater than 1 that indicates cluster membership for a specific number of clusters. For instance if you type 3 into the box indicating number of clusters, SPSS will print out a 3-cluster solution.

**Range of solutions:** If you wish to see several possible solutions, type the value for the smallest number of clusters in the first box and the value for largest number of clusters you wish to see in the second box. For instance if you type in 3 and 5, SPSS would show case membership for a 3-cluster, a 4-cluster and a 5-cluster solution.

#### Button → Plots ...

An icicle plot of the entire clustering process is included by default. The Dendrogram provides information similar to that covered in the icicle plot but features, in addition, a relative measure of the magnitude of differences between variables or clusters at each step of the process.

#### Button → Method ...



**Cluster Method:** Choose the procedure for combining clusters. The default procedure is called the between-group linkage. SPSS computes the smallest average distance between all group pairs and combines the two groups that are closest. The procedure begins with as many clusters as there are cases (here: 21). At step one, the two cases with the smallest distance between them are clustered. Then SPSS computes distances once more and combines the two that are next closest. After the second step you will have either 18 individual cases and one cluster of 3 cases, or 17 individual cases and two clusters of two cases each. The process continues until all cases are grouped into one large cluster.

**Measure:** Indicate what method is used for distance measuring, the default is Squared Euclidean distance.

**Transform values:** In the used data file vcr.sav, most measures are rated on a 5-point scale, but the listed prices fluctuate between \$200 and \$525, and events, days and extras are simply the actual numbers associated with those variables. The solution suggested by SPSS is to standardize all variables – for example, change each variable to a z-score (with a mean of 0 and a standard deviation of 1). This will give each variable equal metrics, but will give them equal weight as well.

#### Button → Save ...

This procedure deals with saving new variables. If you choose to save new variables, they will appear in the form of a new variable in the last column of you data file and simply include a different coded number for each case. There are three options:

**None:** The default.

**Single Solution:** If you indicate 3, then each case will be coded either 1, 2, or 3.

**Range of Solutions:** If you indicate from 3 to 5, three new variables will be created. One variable that codes cases 1, 2 and 3; a second that codes cases 1, 2, 3 and 4; and a third that codes cases 1, 2, 3, 4 and 5.

You can analyze raw variables or you can choose from a variety of standardizing transformations.

Distance or similarity measures are generated by the Proximities procedure.

## Output of running hierarchical clustering analysis

We performed a hierarchical cluster analysis, selecting all the variables except *brand* in the Variable(s) box and we labeled the cases by *brand*. We further requested the **Dendrogram** in the output. We changed all variables to **z-scores** to yield equal metrics and equal weighting, selected the **Squared Euclidean distance** (the default) method of determining distance between clusters and the **Furthest neighbour** method for clustering, and saved a **3-cluster solution as a new variable**.

The Agglomeration Schedule and the Dendrogram will be given next. An icicle plot displays the information of the Agglomeration Schedule graphically, but won't be given here.

## Cluster Complete Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
<b>1</b>	<b>1</b>	<b>3</b>	<b>.002</b>	<b>0</b>	<b>0</b>	<b>17</b>
2	13	18	2,708	0	0	9
3	12	17	4,979	0	0	14
4	20	21	5,014	0	0	7
5	11	14	8,509	0	0	10
6	5	8	11,725	0	0	8
7	19	20	11,871	0	4	14
8	2	5	13,174	0	6	13
9	13	15	14,317	2	0	12
<b>10</b>	<b>9</b>	<b>11</b>	<b>19,833</b>	<b>0</b>	<b>5</b>	<b>15</b>
11	6	7	22,901	0	0	15
12	10	13	23,880	0	9	16
13	2	4	28,378	8	0	17
<b>14</b>	<b>12</b>	<b>19</b>	<b>31,667</b>	<b>3</b>	<b>7</b>	<b>16</b>
15	6	9	40,470	11	10	18
16	10	12	44,624	12	14	19
17	1	2	47,720	1	13	20
18	6	16	49,963	15	0	19
19	6	10	64,785	18	16	20
20	1	6	115,781	17	19	0

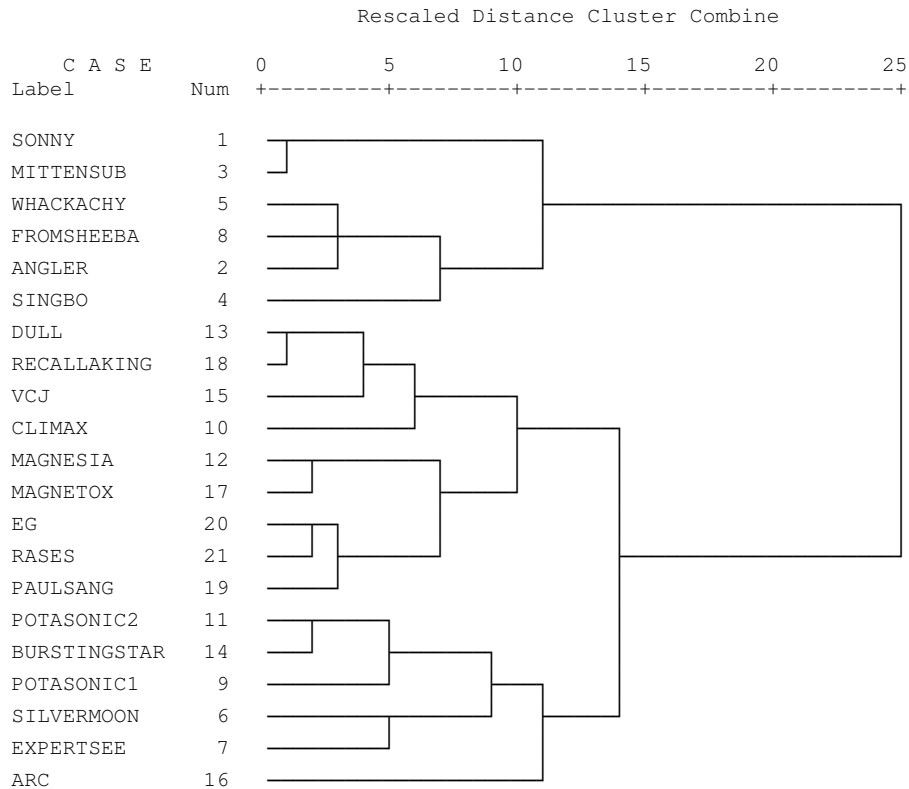
The procedure followed by cluster analysis at Stage 1 is to cluster the two cases that have the smallest squared Euclidean distance between them. Then SPSS will recompute the distance measures between all single cases and clusters (there is only one cluster of two cases after the first step). Next, the 2 cases (or clusters) with the smallest distance will be combined, yielding either 2 clusters of 2 cases (with 17 cases unclustered) or one cluster of 3 (with 18 cases unclustered). This process continues until all cases are clustered into a single group. To clarify, we will explain Stages 1, 10, and 14.

At **Stage 1**, Case 1 is clustered with Case 3. The squared Euclidean distance between these two cases is .002. Neither variable has been previously clustered (the two zeros under Cluster 1 and Cluster 2), and the next stage (when the cluster containing Case 1 combines with another case) is Stage 17. (Note that at Stage 17, Case 2 joins the Case-1 cluster.)

At **Stage 10**, Case 9 joins the Case-11 cluster (Case 11 was previously clustered with Case 14 back in Stage 5, thus creating a cluster of 3 cases: Cases 9, 11, and 14). The squared Euclidean distance between Case 9 and Case-11 cluster is 19.833. Case 9 has not been previously clustered (the zero under Cluster 1), and Case 11 was previously clustered at Stage 5. The next stage (when the cluster containing Case 9 clusters) is Stage 15 (when it combines with the Case-6 cluster).

At **Stage 14**, the clusters containing Cases 12 and 19 are joined, Case 12 has been previously clustered with Case 17, and Case 19 had been previously clustered with Cases 20 and 21, thus forming a cluster of 5 cases (Cases 12, 17, 19, 20, 21). The squared Euclidean distance between the two joined clusters is 31.667. Case 12 was previously joined at Stage 3 with Case 17. Case 19 was previously joined at Stage 7 with the Case-20 cluster. The next stage when the Case-12 cluster will combine with another case/cluster is Stage 16 (when it joins with the Case-10 cluster).

**Dendrogram using Complete Linkage**



The branching-type nature of the Dendrogram allows you to trace backward or forward to any individual case or cluster at any level. It, in addition, gives an idea of how great the distance was between cases or groups that are clustered in a particular step, using a 0 to 25 scale along the top of the chart. While it is difficult to interpret distance in the early clustering phases (the extreme left of the graph), as you move to the right relative distance become more apparent. The bigger the distances before two clusters are joined, the bigger the differences in these clusters. To find a membership of a particular cluster simply trace backwards down the branches to the name.

The **new variable clu3\_1**, regarding the 3-cluster solution, is now visible in the data file, see following figure. You can see that, for example, Brand 1 is clustered into cluster 1, and Brand 6 into cluster 2.

	extras3	clu3_1
1	13	1
2	12	1
3	13	1
4	7	1
5	11	1
6	8	2
7	8	2
8	9	1
9	8	2
10	4	2