Contents

The hierarchical clustering 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.

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.

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.

0.1 Linkage methods

- Single linkage (minimum distance)
- Complete linkage (maximum distance)
- Average linkage

0.1.1 Ward's method

- Compute sum of squared distances within clusters
- Aggregate clusters with the minimum increase in the overall sum of squares

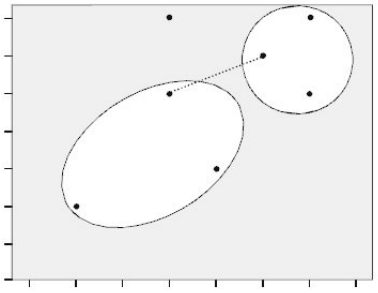
0.1.2 Centroid method

The distance between two clusters is defined as the difference between the centroids (cluster averages)

- A commonly used approach in hierarchical clustering is ***Wards linkage 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 data set does not include outliers, you should always use Wards method.

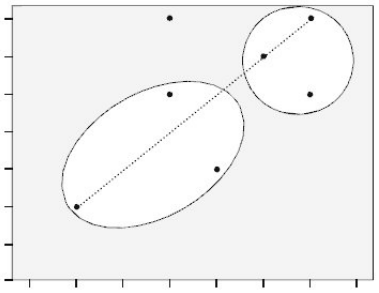
  We will use the Ward's linkage method for laboratory exercises.

- Other most popular agglomerative clustering procedures include the following:

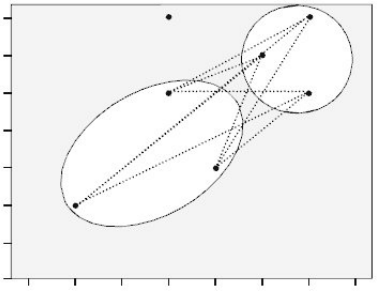


Single linkage (nearest neighbor) : The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.

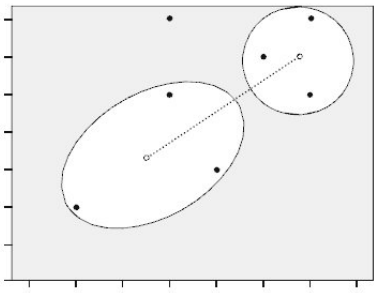
Complete linkage (furthest neighbor) : The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.



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. The distance between the two clusters equals the distance between the two centroids.



Each of these linkage algorithms can yield totally different results when used on the same data set, as each has its specific properties. As 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. Generally, single linkage is considered the most versatile algorithm.

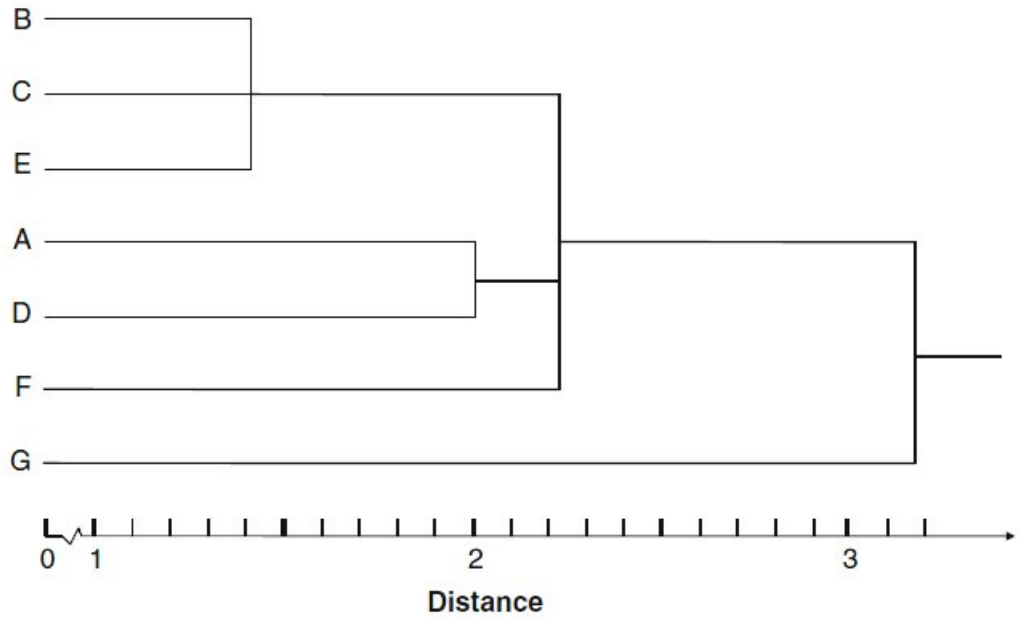
Conversely, the complete linkage method is strongly affected by outliers, as it is based on maximum distances. Clusters produced by this method are likely to be rather compact and tightly clustered. The average linkage and centroid algorithms tend to produce clusters with rather low within-cluster variance and similar sizes. However, both procedures are affected by outliers, though not as much as complete linkage.

An understanding of linkage method's other than than Ward method will be expected in the end of year examination.

- A common way to visualize the cluster analysiss progress is by drawing a dendrogram, which displays the distance level at which there was a combination of objects and clusters. Here is an example of a dendrogram (which corresponds to the example in the next section of material.

- An important question is how to decide on the number of clusters to retain from the data. 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 factor analysiss scree plot, 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, of course. For this purpose, we can make use of the dendrogram.

- In constructing the dendrogram, SPSS rescales the distances to a range of 025; that is, 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. Despite this, 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]).

- The clustering algorithm is based on a distance measure that gives the best results if all variables are independent, continuous variables have a normal distribution (or categorical variables have a multinomial distribution). This is seldom the case in practice, but the algorithm is thought to behave reasonably well when the assumptions are not met.
- Because cluster analysis does not involve hypothesis testing and calculation of observed significance levels, other than for descriptive follow-up, it's perfectly acceptable to cluster data that may not meet the assumptions for best performance.
- The final outcome may depend on the order of the cases in the file. To minimize the effect, arrange the cases in random order. Sort them by the last digit of their ID numbers or something similar.