Cluster Analysis

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Outline I

Outline II

Problems/Questions

Cluster analysis was developed in taxonomy. The aim was originally to get away from the high degree of subjectivity when single taxonomists performed a grouping.

- Clustering is used to build groups of genes with related expression patterns (co-expressed genes): "In analyzing DNA microarray gene-expression data, a major role has been played by various cluster-analysis techniques, most notably by hierarchical clustering, K-means clustering and self-organizing maps. These clustering techniques contribute significantly to our understanding of the underlying biological phenomena." Genome Biology 2002, 3(2):research0009.1–0009.8.

Problems/Questions, cont.

- In plant and animal ecology, clustering is used to describe and to make spatial and temporal comparisons of communities of organisms in heterogeneous environments; in plant systematics to generate artificial phylogenies or clusters of organisms at the species, genus or higher level that share a number of attributes.
- CA might be used to classify regions based on vegetation communities or abundances of species.
- We may find out various stakeholders regarding a national park via conducting a survey; questioning farmers, visitors, inhabitants of the municipalities ...

Idea of Cluster analysis

- Clusters are formed numerically – on the basis of distance measures. Interpretation of clusters is the final step.
- The resulting clusters should exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity.
- The Interpretation of clusters has to take into account (needs to be consistent with) the mathematical procedures.

Idea of Cluster analysis, Cont.

- Comparing clusters (of cases) with respect to additional variables (i.e. variables, which have not been considered in the formation of the clusters) can help in the interpretation of clusters (subsequent to CA).
- CA itself is usually not combined with the calculation of statistical significance.
- CA does – likewise to Factorial Analysis - serve data reduction purposes.

Cluster analysis of cases

- Cluster analysis evaluates the similarity of cases (e.g. persons, animals, areas, other entities) with respect to a defined set of variables.
- Cases are grouped into clusters on the basis of their similarities. Similar cases shall be assigned to the same cluster. Dissimilar cases shall be assigned to different clusters.
- The number of clusters to be formed can be defined in advance or certain criteria are defined and applied to the data.

Cluster analysis of variables

- When variables are clustered, the similarity of the variables is evaluated with respect to the similarity of the values, which a predefined set of persons have on these variables.
- Similar variables shall be assigned to the same cluster of variables. Dissimilar variables shall be assigned to different clusters.
- The number of clusters to be formed can be defined in advance or certain criteria are defined and applied to the data.

Clustering cases or variables?

- Whether cases or variables should be clustered depends on the research question you have.
- Mathematically there exists no fundamental difference between clustering cases or variables.
- Clustering variables starts with the Transposed matrix, as compared to the matrix you start with when clustering cases.

Similarities of cases wrt. their values of certain variables (clustering cases) versus similarities of variables wrt. the values of certain cases on these variables (clustering variables).

Distance measures in CA I

- The definition of the distance measure, and deciding whether to use standardized or raw data are fundamental steps in CA.
- Measures of distance (dissimilarity versus similarity/proximity) $\rightarrow$ Recursively most similar clusters are unified.
- In most CA approaches, the first round of a sequential clustering procedure assumes no pre-existing clusters of cases. Instead, in the first round each case is regarded as a cluster $\rightarrow$ **agglomerative methods**.

Distance measures in CA II

- Accordingly in the first round of a sequential CA procedure the distance is measured between all cases, that is, if n cases are included, the number of distances to be calculated equals $\frac{n \cdot (n-1)}{2}$.
- The scale level of your data can limit the number of distance measures which make sense.
- If a variable is dichotomous two cases can only be identical or different from each other.

Distance measures in CA III

- If variables are ordinal scaled (ranks), then the distances between values are difficult to interpret.
  $\rightarrow$ It is problematic to use distance measures, which are appropriate for interval scale data, for ordinal scaled data.
  $\rightarrow$ A possibility is splitting ranks using the median, and thus recode the variables (e.g. 0 = below median; 1 = above median) to subsequently apply a distance measure for dichotomous variables.

Distance measures in CA IV

- If a variable is nominal scaled with more than two levels (e.g. nationality; region of residence), then a dummy coding of the variables can transform them into dichotomous variables.
- If your variables are interval scaled (quantitative variables), then distances can be properly interpreted.

General properties of a distance measure

- $d(x, y) \geq 0$, the distance is never negative
- $d(x, y) = 0$, if $x = y$
- $d(x, y) = d(y, x)$, symmetric
- $d(x, y) \leq d(x, z) + d(z, y)$

Distance measures for interval scale data

Cases/Variables	V1	V2	V3	V4	V5	V6	V7
1	5	2	3	0	1	0	1
2	4	4	3	3	1	1	1
3	10	7	8	5	6	5	6

Different aspects of similarity (distance) can be focused:

- Cases 1 and 2 are similar regarding the absolute values of the variables $\rightarrow$ most distance measures (e.g. Euclidean distance) focus this aspect
- Cases 1 and 3 ...

Distance measures for interval scale data

Cases/Variables	V1	V2	V3	V4	V5	V6	V7
1	5	2	3	0	1	0	1
2	4	4	3	3	1	1	1
3	10	7	8	5	6	5	6

- Cases 1 and 2
- Cases 1 and 3 are similar with respect to the profile (increase and decrease/relative values) over the variables.
  $\rightarrow$ To focus on this aspect, **product moment correlation** can be selected as similarity measure!

Distance measures for interval scale data

Cases/Variables	V1	V2	V3	V4	V5	V6	V7
1	5	2	3	0	1	0	1
2	4	4	3	3	1	1	1
3	10	7	8	5	6	5	6

- Cases 1 and 2
- Cases 1 and 3 are similar with respect to the profile over the variables.
  $\rightarrow$ After standardizing the variables (e.g. z-transforming) conventional distance measures (e.g. Euclidean distance) also result in a focus on this latter aspect.

Euclidian distance

- The Euclidean Distance between two cases A and B is the "straight line" between the two cases. Assume $A(x_A, y_A), B(x_B, y_B)$, then $d_{Euklidean}(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$.
- The value of the Euclidean distance depends on the scale of the variables. $\rightarrow$ Standardize the variables!
- Generally, two cases having the variable vectors x and y: $d(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$, where k is the number of variables.

Minkowski Metrik

- General basis of both, the City-Block-Distance and the Euclidean Distance (ordinary distance, Pythagorean metric) are the so called Minkowski-Metrics, corresponding to the formula: $d(x, y) = (\sum_{i=1}^{k} |x_i - y_i|^r)^{\frac{1}{r}}$, where r is called the Minkowski constant.
- In the City-Block Metric $r = 1$, in Euclidean Distance $r = 2$. These distance measure can be calculated for any number of variables (dimensions).

Minkowski Metrik, Cont.

- High values of r increase the weight of large distances relative to small ones.
- Dominance (Supremum) metric: $r \to \infty$, thus only the biggest difference matters
- For applying Minkowski-Metrics the scales of the variables should be identical. Else, a standardization (e.g. z-transformation of all variables) must be accomplished.

Cosine

- The distance between two cases x and y is defined as:

$$d(x, y) = \frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

- The direction of the variable vectors is decisive and not their length. The "angle" between the two variable vectors is computed.

Pearson Correlation

- Correlation coefficient between x and y

$$r_{x,y} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{k}(x_i - \bar{x})^2 \sum_{i=1}^{k}(y_i - \bar{y})^2]^{\frac{1}{2}}}$$

- Measure of Similarity
- Note: Clustering variables using Pearson correlation as distance measure is similar to Factorial Analysis as both are closely related to the correlation matrix of the involved variables and both aim to identify similarities between these variables.

Distance measures for binary variables

Cases/variables	V1	V2	V3	V4	V5	V6	V7
1	1	1	1	0	1	0	1
2	0	1	0	0	0	1	1
3	1	1	0	0	0	0	0

Even for binary data manifold distance (similarity) measures exist, e.g.:

1. Simple Matching Coefficient (SMC) = matches/number of paired variables.
   $\rightarrow$ e.g. SMC (case 1, case 2) = 3/7; SMC (case 2, case 3) = 4/7.

Distance measures for binary variables

Cases/variables	V1	V2	V3	V4	V5	V6	V7
1	1	1	1	0	1	0	1
2	0	1	0	0	0	1	1
3	1	1	0	0	0	0	0

1. Simple Matching Coefficient.
2. Jaccard Matching coefficient = matches with characteristic present (=1)/ number of paired variables where characteristic is given (=1) at least once.
   $\rightarrow$ e.g. JMC (case 1, case 2) = 2/6; JMC (case 2, case 3) = 1/4

Distance measures for binary variables

Cases/variables	V1	V2	V3	V4	V5	V6	V7
1	1	1	1	0	1	0	1
2	0	1	0	0	0	1	1
3	1	1	0	0	0	0	0

1. Simple Matching Coefficient.
2. Jaccard Matching coefficient.
3. Phi-coefficient $\rightarrow$ Product-moment correlation formula applied to binary data that is coded 0 and 1.

In these three methods, the distance measure is defined by: d = (1 - similarity measure).

Comments regarding distance measures

- Specific distance measures for nominal scaled variables that are not binary as well as distance measures for ordinal scaled variables are available.
- Act with caution if variables with different level of measurements are used combined!
- Analogous to the distances between cases, distances between variables can also be calculated.

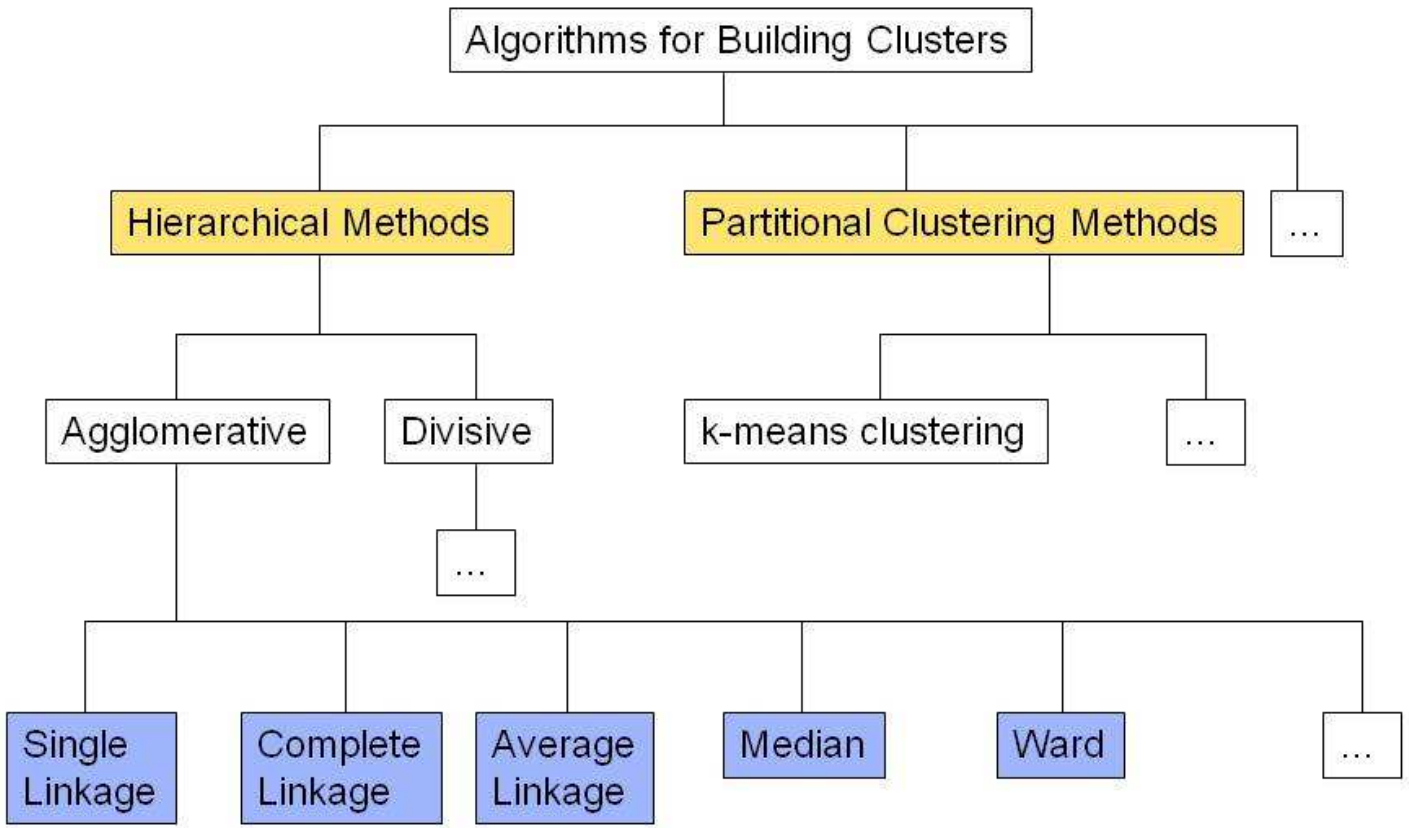
Formation of groups (clusters) in CA

- Different methods can be used to form clusters.
- **Hierarchical Agglomerative CA** methods start with the finest partitioning (i.e. each case [variable] forms one cluster) and sequentially reduce the number of clusters by 1 through unifying two clusters.
  Divisive methods which start with one super cluster are very uncommon and not touched here.

Formation of clusters in CA, Cont.

- In **hierarchical CA methods**, clusters are sequentially enlarged. Once included an element is included in a cluster it will not be excluded from the cluster.
- In **non-hierarchical CA methods**, elements can sequentially be included and excluded into different clusters in order to identify the best cluster structure.
  $\rightarrow$ Thus, they are more flexible and complex.
  $\rightarrow$ The k-means method is the most frequently applied non-hierarchical CA method.

Algorithms for establishing clusters



Hierarchical agglomerative CA methods

- Different hierarchical CA methods use different criteria for selecting which two clusters will be unified in the next step.
- Hierarchical Methods are for example Single Linkage, Complete Linkage, Average Linkage, [weighted/unweighted] Centroid-method, and the Ward-method.

Hierarchical agglomerative CA methods, Cont.

- Even if the same distance measure is used, these methods result in different distances between clusters $\rightarrow$ i.e. different distance matrices (But only if at least one of the clusters contains more than one case).
- Ward method is somewhat different approach as it uses a sums of square criterion and unifies the two clusters which results in the smallest increase of error variance.

The procedure of hierarchical methods

1. Start with finest grouping, i.e. each case is a cluster.
2. Calculation of distance matrix.
3. The two cluster with the lowest distance are identified and unified to a common cluster.
4. A new distance matrix for the reduced number of clusters is calculated (where the two formerly separated cluster are now considered as one cluster).
5. The two cluster with the lowest distance in the reduced distance matrix are unified.

The procedure of hierarchical methods, Cont.

... Recursion to step 4, ... etc. - until all cases are unified in a single super cluster or some predefined criterion for stopping the agglomeration schedule is defined.

Single Linkage

1. Single Linkage calculates the distance $D(R; P\&Q)$ between a (divisible) cluster containing the two cases P, Q, and the indivisible cluster R as:

$$D(R; P\&Q) = \min\{D(R; P), D(R; Q)\}$$

Here, the smallest distance between a cluster $P\&Q$ and the cluster R determines the distance between the two clusters.

Single Linkage, Cont.

The method is also known as Nearest Neighbor as those two clusters, which have the nearest neighbor (i.e. where the pair with the smallest distance between two objects of the two clusters exists) are unified.

This results in a tendency towards the formation of large clusters.

Complete Linkage

2. Complete Linkage calculates the distance $D(R; P\&Q)$ between a (divisible) cluster containing the two cases P, Q, and the indivisible cluster R as:

$$D(R; P\&Q) = \max\{D(R; P), D(R; Q)\}$$

Here, the largest distance between a cluster $P\&Q$ and the cluster R determines the distance between the two clusters.

Complete Linkage, Cont.

The method is also known as Furthest Neighbor as the two clusters where the furthest neighbor (i.e. The pair with the largest distance between two objects of the two clusters) is least distant are unified.

This results in a tendency towards the formation of small clusters.

Average Linkage between groups

3. Average Linkage between groups calculates the distance $D(R; P\&Q)$ between a (divisible) cluster containing the two cases P, Q, and the indivisible cluster R as:

$$D(R; P\&Q) = \text{mean}(D(R; P), D(R; Q))$$

Here, the average of the pair-wise distances between all the pairs formed by objects of both clusters determines the effective distance between the two clusters.

Average Linkage between groups, Cont.

3. Average Linkage:

$$D(R; P\&Q) = \text{mean}(D(R; P), D(R; Q))$$

Here, the average of the pair-wise distances between all the pairs formed by objects of both clusters determines the effective distance between the two clusters.
$\rightarrow$ The two clusters where the average between group distance is smallest are unified.
$\rightarrow$ This method is frequently used.

Average Linkage within groups

4. Average Linkage within groups calculates the mean distance $D(R; P\&Q)$ between all cases in the cluster to be formed out of the divisible cluster and the indivisible cluster R:

$$D(R; P\&Q) = \text{mean}(D(R; P), D(R; Q), D(P; Q))$$

Here, the average of the pair-wise distances between all the objects in the new cluster is calculated for each possible cluster fusion. The two clusters having the lowest average within cluster distance are unified.

Cluster centroids

- For calculating the distances between two large clusters, the 4 previous methods function analogous to our example of an indivisible cluster (R) and a composed cluster with two cases (P,Q).
- In the four methods described so far, individual objects of the (non-elementary) composed clusters are considered for calculating the distance between two clusters.
  The subsequent methods focuses on the two cluster centroids.

Unweighted Centroid Method

5. The unweighted Centroid Method (= Median Method) calculates the distance between clusters as the distance between the centroids of the clusters.

- In principle, the position of the centroid of a Cluster is defined by the average value of the objects forming the cluster of each of the variables that are considered in the CA. In the case of an elementary cluster with only one object, this object represents the centroid.

Unweighted Centroid Method, Cont.

- However, in the unweighted Centroid Method (= Median Method) when calculating the distances between two clusters only the two centroids are considered instead of the single objects within the two clusters.
  $\rightarrow$ The centroid of the cluster resulting from the fusion of two clusters results from the means of the two clusters.
  $\rightarrow$ Large and small clusters are not weighted differentially when they are unified.

Weighted Centroid Method

In the unweighted centroid method the size of the two sub-clusters that are unified is ignored.
$\rightarrow$ The resulting centroid is usually different from the centroid of all elementary objects (i.e. cases) contained in the unified cluster (it is rather the centroid of the two cluster centroids)

$\Downarrow$

The (weighted) Centroid Method considers the differences in the size (number of objects) of the two original sub-clusters.

Weighted Centroid Method, Cont.

Thus in the case of the fusion of a small and a large cluster, the centroid of the resulting cluster is closer to the centroid of the large cluster than to the centroid of the small cluster.

Ward Method (Min. Variance Method)

7. The idea has much in common with an ANOVA. The two clusters that are connected to the smallest increase in the error sums of squares (ESS) are sequentially unified.

Let X_{ijk} denote the value for variable k in observation j belonging to cluster i:

$$\begin{aligned} ESS(X) &= \sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{i\cdot k})^2 \\ &= \sum_{clusters} \sum_{cases} (X_{ij} - centroid_i)^2 \end{aligned}$$

Ward Procedure I

1. At the beginning each case is a cluster.
2. In the first step of the algorithm, $n - 1$ clusters are formed, one of size two and the remaining of size 1. The error sum of squares is computed. The pair of sample units that yield the smallest *ESS* will form the first cluster.

Ward Procedure II

3. Then, in the second step of the algorithm, $n-2$ clusters are formed from that $n-1$ clusters defined in step 2. These may include two clusters of size 2, or a single cluster of size 3 including the two items clustered in step 1. Again, the value of ESS is minimized.
4. Thus, at each step of the algorithm clusters or observations are combined in such a way as to minimize the results of error from the squares.
5. The algorithm stops when all sample units are combined into a single large cluster of size n.

Ward method, Cont.

- The Ward method is frequently applied.
- Homogeneous clusters.
- It has a tendency to generate clusters of similar size (i.e. with similar numbers of cases).
- However, the Ward method's tendency towards equally large clusters can be an advantage (e.g. if homogenous cluster sizes are aimed at) as well as a disadvantage (e.g. if unbalanced cluster sizes do better reflect the reality). In the latter case, Average Linkage or the (weighted) Centroid method could produce better results.

Partitional clustering methods

- A partitioning in G clusters is given (A predefined clustering solution can be used as starting point).
- Number of clusters remains constant.
- The partitioning is suboptimal and the algorithm tries to improve it.
- Iteratively rearrangements improve the partitioning.
- The methods differ in the criteria to measure this improvement and the rules for the rearrangements.

k-means clustering

- Follows the idea of an ANOVA. The clusters are established in order to maximize the F-statistic: The ratio of the variance between clusters and the variance inside the clusters is maximized.
- The main advantages of this algorithm are its simplicity and speed which allows it to run on large data sets.

k-means clustering-algorithm

- Randomly generate G clusters/centroids.
- Compute the centroids for each cluster.
- Calculate for each case the distance to the centroid of each cluster.
- Move the case into the cluster with the smallest distance.
- Repeat the above two steps till no change in the cluster appears.

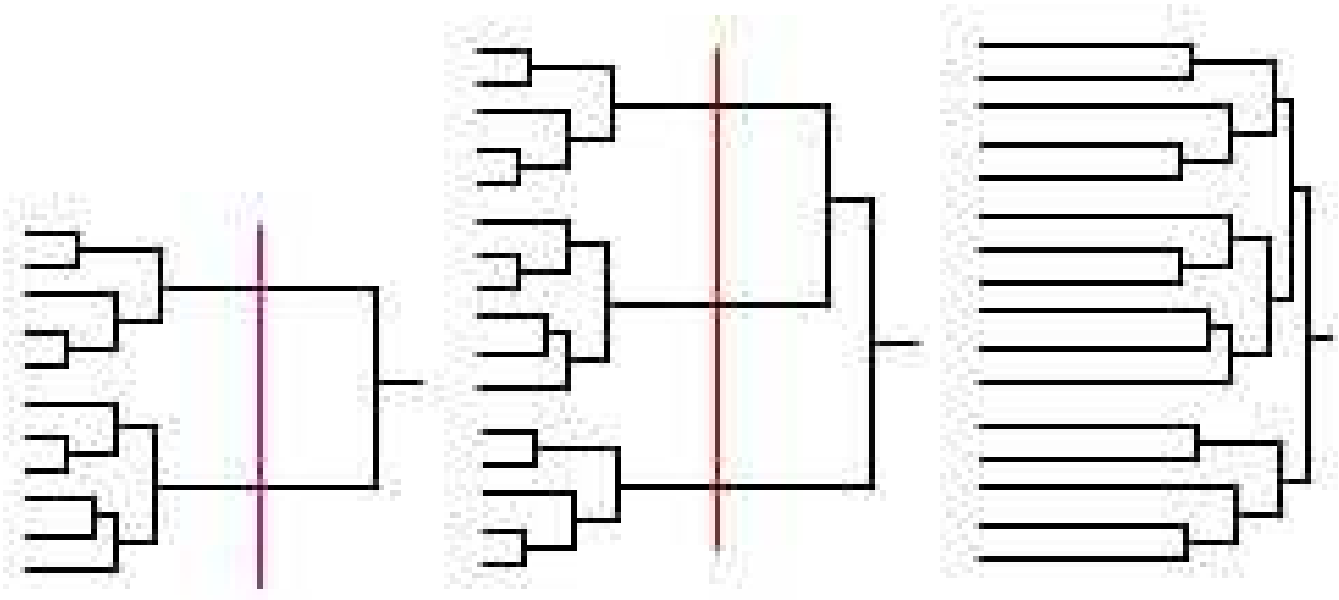
Dendrogram

The results of a CA can be depicted in a Dendrogram. It shows the sequence in which the clusters have been formed and the increases of within cluster distances that are connected with the formation of ever larger clusters (i.e., the increase of error variance in the case of the Ward Method)

Dendrogram, Cont.

- Together with the classification of the objects, the Dendrogram is the most interesting output of a CA.
- It can be used to judge the homogeneity of the clusters.
- It can also be used to define the number of clusters that should be used for the final classification of the objects

Dendrogram, Cont.



Further Analysis of the obtained clusters I

- If the cases in the data set are assigned to clusters (clustering of cases), then the number of the cluster in which each case is included can be saved as a new variable.
- The number of clusters to be formed can be predefined or a range of solutions can be generated.
- Subsequent analyzes comparing the different clusters with respect to variables of interest becomes possible.

Further Analysis of the obtained clusters II

- ANOVA.
- Discriminant Analysis.
- If clusters differ significantly with respect to variables of interest that are rather independent of the variables which have been used for clustering the cases, then the meaningfulness of the clusters and the usefulness of the clustering procedure becomes evident.