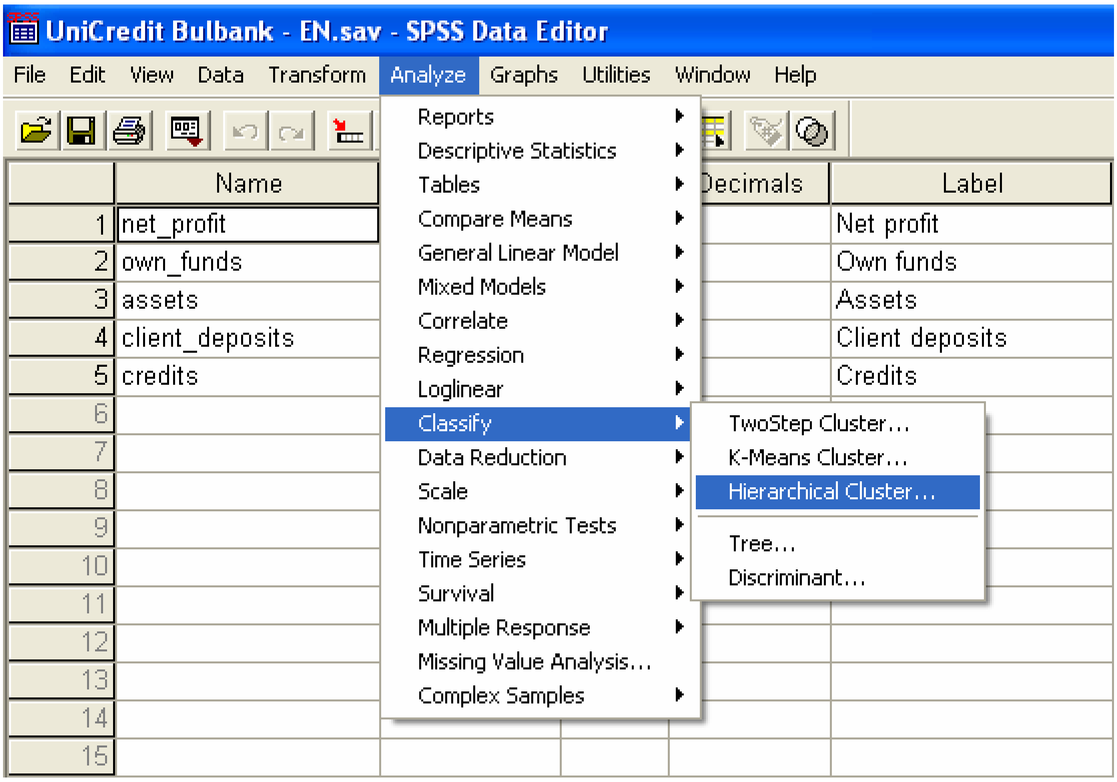
# Cluster analysis with SPSS: Hierarchical Cluster Analysis

From the main menu consecutively click **Analyze** → **Classify** →**Hierarchical Cluster**.

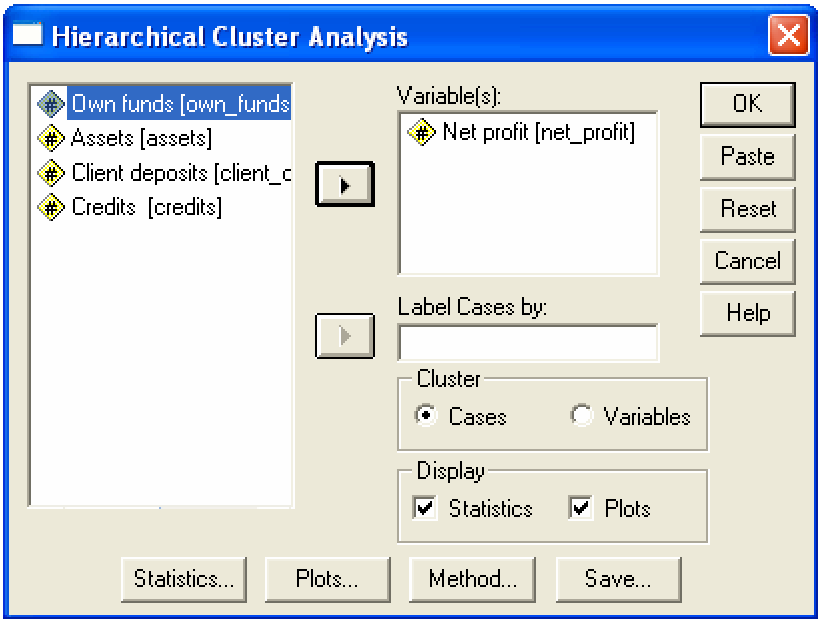


UniCredit Bulbank - EN.sav - SPSS Data Editor

|  | Name | Decimals | Label |
|---|---|---|---|
| 1 | net_profit |  | Net profit |
| 2 | own_funds |  | Own funds |
| 3 | assets |  | Assets |
| 4 | client_deposits |  | Client deposits |
| 5 | credits |  | Credits |

**Figure 1.**

The following dialog window appears:



Hierarchical Cluster Analysis

Own funds [own_funds
Assets [assets]
Client deposits [client_c
Credits [credits]

Variable(s):
Net profit [net_profit]

Label Cases by:

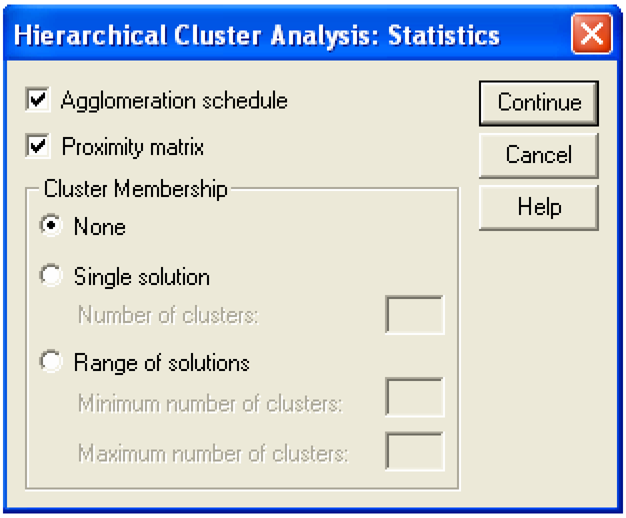
Cluster
● Cases ○ Variables

Display
☑ Statistics ☑ Plots

**Figure 2.**

Select the variables to be analyzed one by one and send them to the **Variables** box. Later actions greatly depend on which type of clustering is chosen here. For this purpose from the **Cluster** box we can choose between **Cases**, where we are performing clustering of the objects and **Variables,** where we are performing clustering of the variables. Next we must set the method for identifying the objects. The **Label Cases by** box is used for entering a string variable which labels the units. If instead of Cases (objects) we set Variables in the Cluster box, then we are required to set the variables in the Variable(s) list, and the Label Cases box is left empty. The default settings for the **Display** box are **Statistics**, for displaying the statistic results from the analysis and **Plots**, for displaying graphs. For both cases it is not necessary to remove the ticks.

Four buttons for entering additional commands are located in the lower part of the dialog window.  We activate the **Statistics** button (figure 2), which is used for defining the statistic results displayed on the screen. Here we can tick **Agglomeration schedule** box for displaying the agglomeration schedule or **Proximity Matrix** for displaying the proximity matrix, which presents the information for the distances between the objects and the clusters. In the first case, the sequence of unification of the objects in clusters is visualized and in the beginning each object is considered as a separate cluster and then the clustering is initiated. Further down in the **Cluster Membership** box we can choose: **None -** if it is not necessary to display the cluster membership of the objects; **Single solution** – when the exact number of the clusters is set and **Range of solutions** – when we set the range of the desired clusters – what is the maximum and minimum number of clusters we want to receive. We proceed by clicking Continue.



Hierarchical Cluster Analysis: Statistics

☑ Agglomeration schedule

☑ Proximity matrix

Cluster Membership

● None

○ Single solution

Number of clusters:

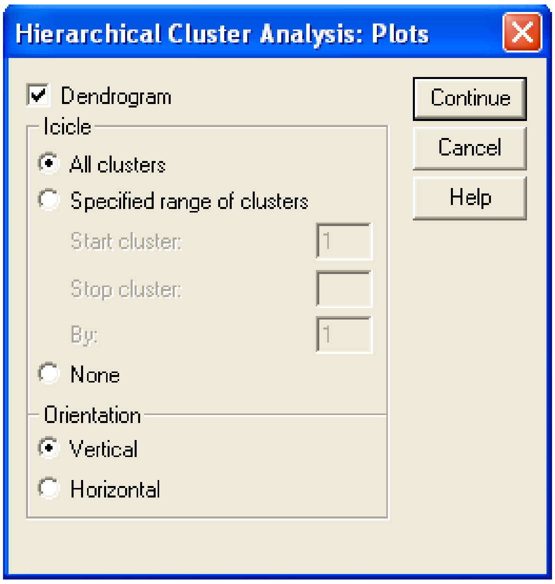
○ Range of solutions

Minimum number of clusters:

Maximum number of clusters:

**Figure 3.**

When we activate the **Plots** button we can select **Dendrogram**, if we want a graphic visualization of the results from the hierarchical clustering. The Dendrogram is a tree graph in which each node represents a stage from the clustering process. It gives additional information about the magnitude of the distance between the two clusters at the moment of unification. The horizontal dotted line of the dendrogram indicates the rescaled distance, in which the clusters are formed. In **Icicle** with **All clusters** selected we set the diagram to include all clusters, using **Specified range of clusters** we can specify the range of the clusters, and with **None** we cancel the icicle. By means of the buttons in the **Orientations** box we can choose the type of the diagram – vertical or horizontal.



Hierarchical Cluster Analysis: Plots

☑ Dendrogram

Icicle

● All clusters

○ Specified range of clusters

Start cluster: 1

Stop cluster:

By: 1

○ None

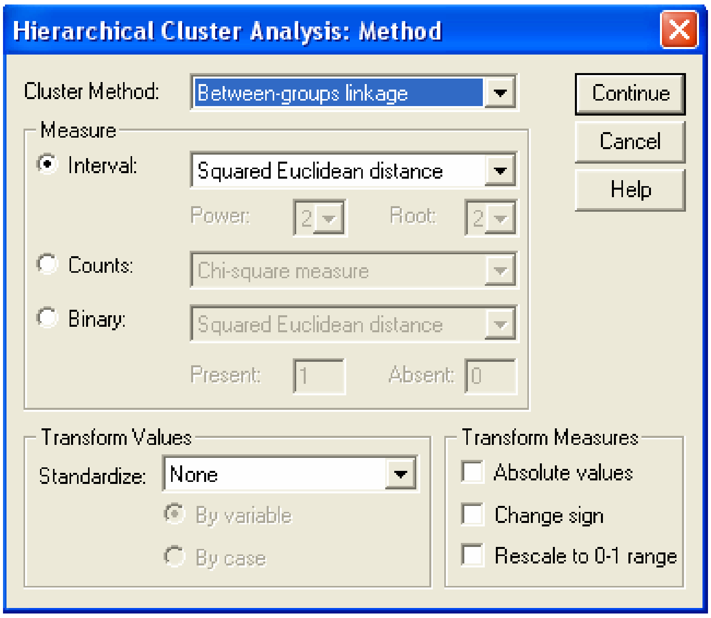
Orientation

● Vertical

○ Horizontal

**Figure 4.**

When we click on **Method,** the following dialog window appears in the **Hierarchical Cluster Analysis** window:



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

● Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-square measure

○ 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 5.**

Firstly, with **Cluster Method** we specify the cluster method which is to be used. With SPSS there are 7 possible methods:

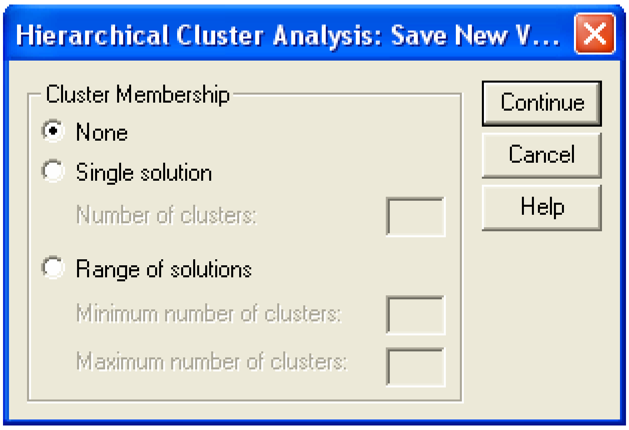
- Between-groups linkage method
- Within-groups linkage method
- Nearest neighbor method
- Furthest neighbor method
- Centroid clustering method
- Median clustering method
- Ward’s method

Each one of these methods leads to different clustering. It cannot be determined which one is the best, but if we are looking for clusters in the form of a „chain” it is advisable to use the methods of Between-groups linkage and Nearest neighbor. When we are looking for clusters in the form of a “cluster”, it is advisable to use the methods of Within-groups linkage and Furthest neighbor. It is important to note that unlike the “cluster” type of clusters, in the “chain” type of clusters the number of objects in the different clusters is considerably different.

In **Measure** box we must determine the convergence measure, i.e. the method for measuring the similarity and divergence between the units.  We choose it according to the measuring scale of the used variables – whether it is interval (**Interval**), categorial (**Counts**) or binary (**Binary**). In other words we determine measures for similarity and divergence for numeric, non-numeric or alternative variables. In the **Transform Values** box using **Standardize** we have to determine the method for standardization of the variables. The different methods were described earlier.

In the **Transform Measures** box using **Absolute Values** we eliminate the direction of the linkage when we use the correlation coefficient when grouping variables. Using **Change Sign** we change the sign before similarity measures turning them into divergence measures and vice versa, and using **Rescale to 0-1 Range** we standardize the similarity and divergence measures in the interval from 0 to 1, when it is necessary.

Using the **Save** button we save the clustering results, i.e. the membership of each object to the corresponding cluster, as a different variable in the data file. Here we can determine the method for saving the cluster membership label for each unit. With **None** we cancel saving cluster membership label for each unit. Using **Single Solution** we can save the cluster membership label for each unit when the number of the clusters is predetermined, and using **Range of Solutions** we save the cluster membership label for each unit when there is a predetermined sequence of cluster solutions.



Cluster Membership

● None

○ Single solution

Number of clusters: □

○ Range of solutions

Minimum number of clusters: □

Maximum number of clusters: □

**Figure 6.**

Let us go back to the example with the basic indices of UniCredit Bulbank and perform hierarchical clustering using Average Linkage Between Groups. In Table 1 a case summary is presented, i.e. valid, missing and total values.

**Table 1.**

**Case Processing Summary(a)**

| Cases | | | | | |
|---|---|---|---|---|---|
| Valid | | Missing | | Total | |
| N | Percent | N | Percent | N | Percent |
| 7 | 100,0 | 0 | ,0 | 7 | 100,0 |

(a) Average Linkage (Between Groups)

In Table 2 (the Proximity Matrix), which is constructed directly with SPSS, is presented. This matrix contains the Squared Euclidean Distances with divergence measure according to the data from the example. For example, the Squared Distance between the first two years is calculated as follows:

$$s_{il}^{E2} = \sum_{j=1}^{p}(X_{ij} - X_{lj})^2 \quad , \qquad i,l = 1,\ldots,n \, .$$

$s_{12}$ = (160,065-68,912)$^2$ + (602,776-490,479)$^2$ + (2 559,476-2 731,686)$^2$ + (1692,270-2021,634)$^2$ + ( 316,380-362,353)$^2$ = 161 169,931.

**Table 2.**

**Proximity Matrix**

| Case | Squared Euclidean Distance | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | ,000 | 161169,931 | 230196,576 | 674488,593 | 3781328,705 | 3633840,451 | 9195794,008 |
| 2 | 161169,931 | ,000 | 42154,582 | 344379,088 | 2653215,196 | 2731459,825 | 7490302,184 |
| 3 | 230196,576 | 42154,582 | ,000 | 152904,699 | 2241708,704 | 2206177,741 | 6768568,041 |
| 4 | 674488,593 | 344379,088 | 152904,699 | ,000 | 1398061,171 | 1244224,157 | 5140974,232 |
| 5 | 3781328,705 | 2653215,196 | 2241708,704 | 1398061,171 | ,000 | 207861,100 | 1260287,862 |
| 6 | 3633840,451 | 2731459,825 | 2206177,741 | 1244224,157 | 207861,100 | ,000 | 1457084,166 |
| 7 | 9195794,008 | 7490302,184 | 6768568,041 | 5140974,232 | 1260287,862 | 1457084,166 | ,000 |

This is a dissimilarity matrix

The Euclidean Distance to a great extent depends on the measure and scale of the different variables. The variable, which is expressed with larger numbers, has more influence in its calculation.

Let us have a detailed look at the hierarchical clustering process using the Average linkage between groups method after we have obtained the distance matrix.

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| 2000 | | 161 169,931 | 230 196,576 | 674 488,593 | 3 781 328,705 | 3 633 840,451 | 9 195 794,008 |
| 2001 | 161 169,931 | | 42 154,582 | 344 379,088 | 2 653 215,196 | 2 731 459,825 | 7 490 302,184 |
| 2002 | 230 196,576 | 42 154,582 | | 152 904,699 | 2 241 708,704 | 2 206 177,741 | 6 768 568,041 |
| 2003 | 674 488,593 | 344 379,088 | 152 904,699 | | 1 398 061,171 | 1 244 224,157 | 5 140 974,232 |
| 2004 | 3 781 328,705 | 2 653 215,196 | 2 241 708,704 | 1 398 061,171 | | 207 861,100 | 1 260 287,862 |
| 2005 | 3 633 840,451 | 2 731 459,825 | 2 206 177,741 | 1 244 224,157 | 207 861,100 | | 1 457 084,166 |
| 2006 | 9 195 794,008 | 7 490 302,184 | 6 768 568,041 | 5 140 974,232 | 1 260 287,862 | 1 457 084,166 | |

At the first stage of clustering the second and the third year are combined because the distance between them is the least $s_{23}$ = 42 154,582. The dimensions of the distance matrix are reduced by 1 and it has the following elements:

| | 2000 | 2001,2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|
| 2000 | | 391 366,507 | 674 488,593 | 3 781 328,705 | 3 633 840,451 | 9 195 794,008 |
| 2001,2002 | 391 366,507 | | 497 283,787 | 4 894 923,9 | 4 937 637,566 | 14 258 870,23 |
| 2003 | 674 488,593 | 497 283,787 | | 1 398 061,171 | 1 244 224,157 | 5 140 974,232 |
| 2004 | 3 781 328,705 | 489 4923,9 | 1 398 061,171 | | 207 861,100 | 1 260 287,862 |
| 2005 | 3 633 840,451 | 4 937 637,566 | 1 244 224,157 | 207 861,100 | | 1 457 084,166 |
| 2006 | 9 195 794,008 | 14 258 870,23 | 5 140 974,232 | 1 260 287,862 | 1 457 084,166 | |

At the next stage the first and the second cluster are combined (2000 and 2001, 2002) because in accordance with theory we obtain the least mean distance: $s_{12}$ =

391 366,507/2 = 195 683,253. The dimensions of the distance matrix are again reduced by 1:

| | 2000,2001,2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|
| 2000,2001,2002 | | 1 171 772,380 | 8 676 252,605 | 8 571 478,017 | 23 454 664,238 |
| 2003 | 1 171 772,380 | | 1 398 061,171 | 1 244 224,157 | 5 140 974,232 |
| 2004 | 8 676 252,605 | 1 398 061,171 | | 207 861,100 | 1 260 287,862 |
| 2005 | 8 571 478,017 | 1 244 224,157 | 207 861,100 | | 1 457 084,166 |
| 2006 | 23 454 664,238 | 5 140 974,232 | 1 260 287,862 | 1 457 084,166 | |

At the third stage, the third and the fourth cluster are combined (2004 and 2005), where: $s_{34}$ = 207 861,100.

| | 2000,2001,2002 | 2003 | 2004,2005 | 2006 |
|---|---|---|---|---|
| 2000,2001,2002 | | 1 171 772,380 | 17 247 730,622 | 23 454 664,238 |
| 2003 | 1 171 772,380 | | 2 642 285,328 | 5 140 974,232 |
| 2004,2005 | 17 247 730,622 | 2 642 285,328 | | 2 717 372,028 |
| 2006 | 23 454 664,238 | 5 140 974,232 | 2 717 372,028 | |

At the next stage we combine the first and the second cluster (2000, 2001, 2002 and year 2003). In this case $s_{12}$ = 1 171 772,380/3 = 390 590,794.

| | 2000,2001,2002,2003 | 2004,2005 | 2006 |
|---|---|---|---|
| 2000,2001,2002,2003 | | 19 890 015,950 | 28 595 638,470 |
| 2004,2005 | 19 890 015,950 | | 2 717 372,028 |
| 2006 | 28 595 638,470 | 2 717 372,028 | |

At the fifth stage we combine the second and the third cluster, i.e. 2004,2005 and 2006, where the mean distance is the least: $s_{23}$ = 2 717 372,028/2 = 1 358 686,014.

| | 2000,2001,2002,2003 | 2004,2005,2006 |
|---|---|---|
| 2000,2001,2002,2003 | | 48 485 654,420 |
| 2004,2005,2006 | 48 485 654,420 | |

At the final stage we combine the two clusters that are left, where the mean distance is: $s_{12}$ = 48 485 654,420/12= 4 040 471,202.

The results from the different stages of the hierarchical clustering in SPSS are summarized and displayed in a table called **Agglomeration Schedule**. In this table we can also see a column with the mean distances calculated so far. In this case the Squared Euclidean Distance is used as a measure.

**Table 3.**

**Agglomeration Schedule**

| Stage | Cluster Combined | | Coefficients | Stage Cluster First Appears | | Next Stage |
|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| 1 | 2 | 3 | 42154,582 | 0 | 0 | 2 |
| 2 | 1 | 2 | 195683,253 | 0 | 1 | 4 |
| 3 | 5 | 6 | 207861,100 | 0 | 0 | 5 |
| 4 | 1 | 4 | 390590,794 | 2 | 0 | 6 |
| 5 | 5 | 7 | 1358686,014 | 3 | 0 | 6 |
| 6 | 1 | 5 | 4040471,201 | 4 | 5 | 0 |

The first **Stage** column of the agglomeration schedule displays the numbers of the different stages and in the last stage all analyzed objects are combined in one cluster. In most cases they are *n*-1. The columns with the common heading **Cluster Combined** show the numbers of the clusters, which are combined at the different stages. For example, at the first stage the second and the third cluster are combined. The **Coefficients** column gives the mean distances (from the theory) for combining the clusters. These coefficients depend on the method which is chosen for forming the cluster.  The indices in this column can be used for approximate rating of the similarity between clusters which are formed at each stage. Large coefficients (for divergence measures) or small coefficients (for similarity measures) indicate that a cluster is relatively heterogeneous and contains units which are considerably different from each other. Coefficients in this column can also serve as approximate orientation for the number of clusters which must be profiled from a practical point of view. For this purpose we can investigate the stage at which a sudden change in coefficients is noticeable. The columns with the common heading **Stage Cluster First Appears** display the stages at which the respective clusters appeared for the first time, and in the **Next Stage** column can be found the stage at which the respective cluster will appear the next time it combines with another cluster. For example, at the first stage, when the second and the third cluster are combined, a new cluster is formed and is assigned number 2. At the second stage, the newly formed cluster 2 is combined with cluster 1 and so on..

The results from the different stages of the hierarchical clustering can also be illustrated with the so called **Icicle Plot**. It can be vertical or horizontal. Figure 7 presents a vertical icicle of the clustering results of UniCredit Bulbank’s years of development.

**Vertical Icicle**

| Number of clusters | Case |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 7 |  | 6 |  | 5 |  | 4 |  | 3 |  | 2 |  | 1 |
| 1 | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 2 | X | X | X | X | X |  | X | X | X | X | X | X | X |
| 3 | X |  | X | X | X |  | X | X | X | X | X | X | X |
| 4 | X |  | X | X | X |  | X |  | X | X | X | X | X |
| 5 | X |  | X |  | X |  | X |  | X | X | X | X | X |
| 6 | X |  | X |  | X |  | X |  | X | X | X |  | X |

**Figure 7.**

Each row in the **Vertical Icicle** corresponds to the number of clusters. In most cases they are *n*-1. A separate column, which is marked with X all the way to the last row, corresponds to each unit. Between the different columns, which correspond to the units, there are other columns, which are marked differently. The icicle is viewed from the bottom upwards. For example, years 2001 and 2002 are combined in cluster 6 – the column between them is marked up to the last row, i.e. there are 6 X signs. The year 2000 is added to 2001 and 2002, thus forming cluster 5. The column between 2000 and 2001 contains five X, etc. When the combinations are relatively large the information in the vertical icicle can prove to be insufficient. In this case we can use horizontal icicle or graphically present only a part of the clusters.

For graphical presentation of the hierarchical clustering results we can use the so called dendrogram.

```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *


 Dendrogram using Average Linkage (Between Groups)

                         Rescaled Distance Cluster Combine

   C A S E    0         5        10        15        20        25
  Label  Num  +---------+---------+---------+---------+---------+

           2   ⇩⇘
           3   ⇩⇳⇩⇩⇩⇘
           1   ⇩⇙   ▭⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
           4   ⇩⇩⇩⇩⇩⇙                                            ⇔
           5   ⇩⇩⇩×⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘                                    ⇔
           6   ⇩⇩⇩⇙            ▭⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇙
           7   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇙
```

**Figure 8.**

The horizontal dotted line in the dendrogram presented in figure 8 shows the rescaled distance at which the clusters are combined. The minimum distance – in this case 42 154,582 corresponds to 1, and the maximum – 4 040 471,202 corresponds to 25. The dendrogram allows the formulation of the following results:

- Years 2000, 2001 and 2002 are combined in a common cluster with relatively small distance, i.e. the cluster is relatively homogenous;
- Year 2003 forms a separate cluster, which is subsequently combined with the cluster of 2000, 2001 and 2002 with relatively the same indexes and almost double loans.
- Years 2004 and 2005 form a separate cluster and are combined with the cluster of 2006, and they are at considerable distance from the others. They have relatively higher values of all basic indexes, especially year 2006.

We notice that the results obtained using Hierarchical Cluster Analysis are identical with those obtained from K-Means Cluster Analysis by changing cluster centers after the joining of each object to a given cluster.

Author: Dessislava Vojnikova,
Plovdiv University,
4th year Bachelor program in Applied Mathematics

Supervisor: Snezhana Gocheva-Ilieva