Cluster analysis with SPSS: K-Means Cluster Analysis

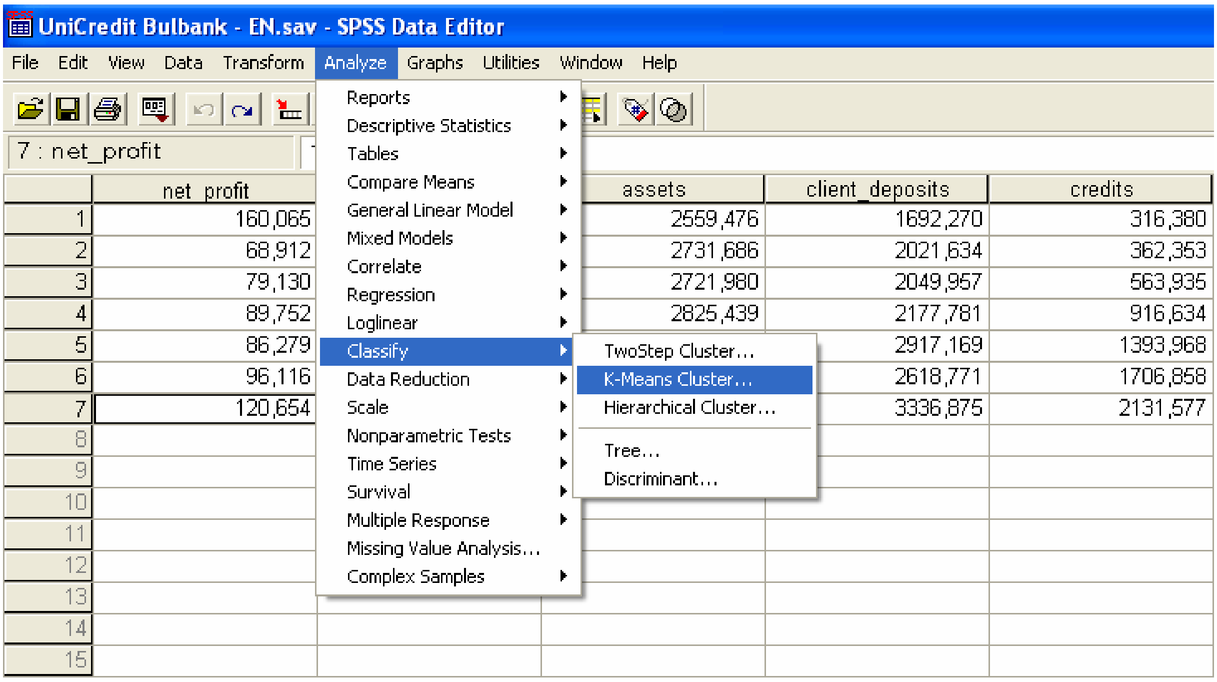
Cluster analysis is a type of data classification carried out by separating the data into groups. The aim of cluster analysis is to categorize n objects in k ($k>1$) groups, called **clusters**, by using p ($p>0$) variables. As with many other types of statistical, cluster analysis has several variants, each with its own clustering procedure.

There are two main sub-divisions of clustering procedures. In the first procedure the number of clusters is pre-defined. This is known as the **K-Means Clustering** method. When the number of the clusters is not predefined we use **Hierarchical Cluster** analysis.

The great variety of clustering procedures results from the metrics which are used between different objects. The most commonly used metrics are the Euclidean metric, Manhattan metric, Chebyshev metric and others. There are also different rules used for creating the clusters. Some allow members to share different clusters, whilst others only allow exclusive membership.

K-Means Cluster Analysis

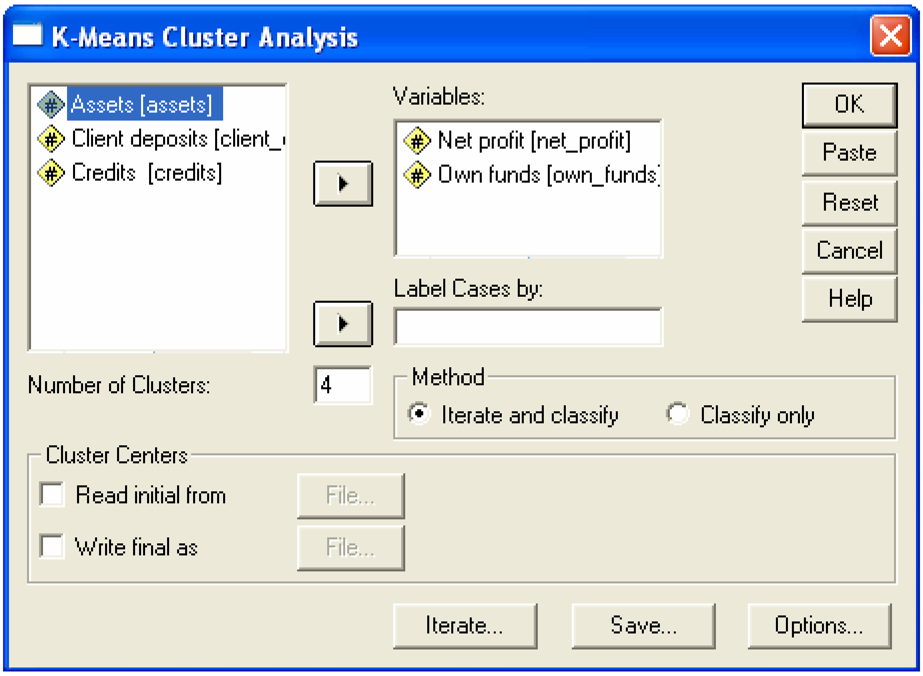
From the main menu of SPSS consecutively click **Analyze**→ **Classify** → **K-Means Cluster**.



	net_profit		assets	client_deposits	credits
1	160,065		2559,476	1692,270	316,380
2	68,912		2731,686	2021,634	362,353
3	79,130		2721,980	2049,957	563,935
4	89,752		2825,439	2177,781	916,634
5	86,279			2917,169	1393,968
6	96,116			2618,771	1706,858
7	120,654			3336,875	2131,577
8					
9					
10					
11					
12					
13					
14					
15					

Figure 1.

Mark the variables on the basis of which clustering will be done and send them to **Variables -** the box for entering the variables. The **Label Cases by** box is used for entering a string variable which marks the units. After that we determine the number of desired clusters in the **Number of Clusters** box. In our case in the **Method** box we mark **Iterate and Classify**. Unlike the alternative method – **Classify only**, which, defines fixed cluster centers, this one defines the successive iterations and determines how the final clustering is to be carried out.



K-Means Cluster Analysis

Assets [assets]
Client deposits [client_
Credits [credits]

Variables:
Net profit [net_profit]
Own funds [own_funds]

Label Cases by:

Number of Clusters: 4

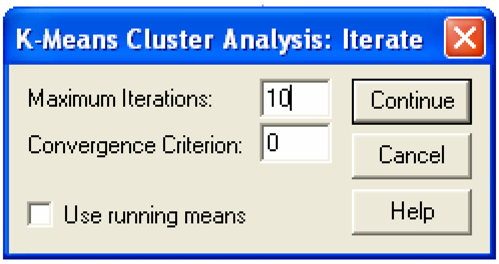
Method
● Iterate and classify ○ Classify only

Cluster Centers
□ Read initial from
□ Write final as

Figure 2.

In the **Cluster Centers** box we specify the file (if there is one), which contains the initial cluster centers and the file (if required), which contains the final cluster centers. In **Read initial from** we specify the file which contains the initial cluster centers, and in **Write final as** we specify the file which contains the final cluster centers.

With the **Iterate** button we can determine the criteria for updating the cluster centers, in **Maximum Iterations** we can determine the maximum number of the iterations (no more than 999), and in **Convergence Criterion** we decide which rule halts the iteration process. By default 10 iterations and convergence criterion 0 are given. Furthermore, it is possible to mark the option **Use running means**. If this is selected, the cluster centers change after the addition of each object. If this option is not selected, cluster centers are calculated after all objects have been allocated to a given cluster. In both cases we receive different results and therefore the way in which the clustering is achieved must be specified. We proceed by clicking on Continue.



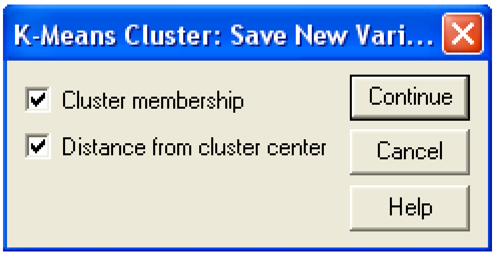
Maximum Iterations: 10

Convergence Criterion: 0

□ Use running means

Figure 3.

Using the **Save** button we can save new variables in a data file which indicates the cluster membership of each object (**Cluster Membership**) and distance from cluster center for each object (**Distance from Cluster Center)**.

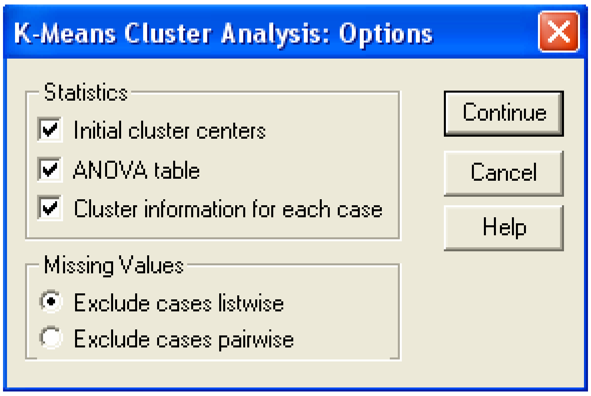


☑ Cluster membership

☑ Distance from cluster center

Fgure 4.

The **Options** button gives the option of displaying additional statistics – initial cluster centers (**Initial cluster centers**), dispersion analysis table (**ANOVA table**) and information for cluster membership of each object (**Cluster information in each case?**). It is desirable that all three options are selected. We obtain the final result by clicking OK.



Statistics

☑ Initial cluster centers

☑ ANOVA table

☑ Cluster information for each case

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

Figure 5.

Let us briefly go through the different stages of K-Means Cluster Analysis using the data from the example with UniCredit Bulbank (Table 1 from the chapter *First Steps in SPSS*). We determine the number of clusters to be 4, and the initial cluster centers are evaluated based on the data. We use squared Euclidean Distance for the divergence measure between units. Also we choose the cluster centers to be calculated after all objects have been assigned to a given cluster, i.e. we don't put a tick in **Use running means** box.

The initial cluster centers are given in Table 1 (**Initial Cluster Centers**). They are vectors with their values based on the five variables, which refer to 2000 (first cluster), 2005 (second cluster), 2006 (third cluster) and 2003 (fourth cluster). These 4 years are at maximum index distance from each other.

Table 1.

Initial Cluster Centers

	Cluster			
	1	2	3	4
Net profit	160,065	96,116	120,654	89,752
Own funds	602,776	609,609	630,781	550,026
Assets	2559,476	3474,829	4346,594	2825,439
Client deposits	1692,270	2618,771	3336,875	2177,781
Loans	316,380	1706,858	2131,577	916,634

In Table 2 we can see the number of the iterations and the changes in the cluster centers. In the first iteration year 2001 joins year 2000 and the cluster center is updated. The year 2004 joins the second cluster – the year 2005, and year 2002 joins the fourth cluster – the year 2003. The third cluster does not change. In the second iteration the process of redistribution of the units stops because there are no changes in the cluster centers.

Table 2.

Iteration History (a)

Iteration	Change in Cluster Centers			
	1	2	3	4
1	200,730	227,959	,000	195,515
2	,000	,000	,000	,000

(a) Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 821,273.

The results are summarized in Table 3, i.e. which cluster each unit belongs to and the new cluster centers. The first cluster is formed by the years 2000 and 2001, the second by 2004 and 2005, the third only by 2006 and the fourth by the years 2002 and 2003.

In Table 4 we can see the final cluster centers, and in Table 5 - the distance between the final cluster centers.

Table 3.

Cluster Membership

Case Number		Cluster	Distance
1:	2000	1	200,730
2:	2001	1	200,730
3:	2002	4	195,515
4:	2003	4	195,515
5:	2004	2	227,959
6:	2005	2	227,959
7:	2006	3	,000

Table 4.

Final Cluster Centers

	Cluster			
	1	2	3	4
Net profit	114,489	91,198	120,654	84,441
Own funds	546,628	591,861	630,781	531,638
Assets	2645,581	3544,763	4346,594	2773,710
Client deposits	1856,952	2767,970	3336,875	2113,869
Loans	339,367	1550,413	2131,577	740,285

Table 5.

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		1762,868	2881,450	494,253
2	1762,868		1143,119	1297,055
3	2881,450	1143,119		2432,395
4	494,253	1297,055	2432,395	

If we compare the results from Table 1 and Table 4 we will see that the cluster center of the third cluster does not change.

Since in our case the groups are formed deliberately in accordance with the distance between them in the multidimensional space, i.e. the condition for randomness of the observations in the different groups is not met, the results from the dispersion analysis are purely descriptive. In other words, we cannot use the significance level (Sign. column in ANOVA Table – dispersion analysis of clustering results) to check the hypothesis about the mean variables. Nevertheless, the differences between the F-ratios (F column in the ANOVA Table) makes it possible to draw general conclusions about the role of the different mean variables in the forming of the clusters.

In Table 6 are given the results from the dispersion analysis. They show that assets have the greatest influence in the forming of the clusters and net profit has the least influence.

Table 6.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Net profit	495,145	3	1419,744	3	,349	,795
Own funds	2878,202	3	2537,200	3	1,134	,460
Assets	842788,443	3	9987,138	3	84,387	,002
Client deposits	634017,636	3	35643,498	3	17,788	,021
Loans	957411,333	3	37401,709	3	25,598	,012

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 7.

Number of Cases in each Cluster

Cluster	1	2,000
	2	2,000
	3	1,000
	4	2,000
Valid		7,000
Missing		,000

Table 7 presents data for the number of units in each cluster as well as their total number and missing units (if there are any).

Now we will present the results from the same clustering procedure with the difference that we choose the cluster centers to be changed after the joining of each object to a given cluster and we select the option **Use running means**.

Table 8.

Iteration History(a)

Iteration	Change in Cluster Centers			
	1	2	3	4
1	215,142	151,973	,000	,000
2	53,786	50,658	,000	,000
3	13,446	16,886	,000	,000
4	3,362	5,629	,000	,000
5	,840	1,876	,000	,000
6	,210	,625	,000	,000
7	,053	,208	,000	,000
8	,013	,069	,000	,000
9	,003	,023	,000	,000
10	,001	,008	,000	,000

(a) Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is ,005. The current iteration is 10. The minimum distance between initial centers is 821,273.

Table 9.

Cluster Membership

Case Number		Cluster	Distance
1:	2000	1	286,856
2:	2001	1	140,021
3:	2002	1	206,434
4:	2003	4	,000
5:	2004	2	227,963
6:	2005	2	227,955
7:	2006	3	,000

Table 10.

Final Cluster Centers

	Cluster			
	1	2	3	4
Net profit	102,702	91,198	120,654	89,752
Own funds	535,501	591,861	630,781	550,026
Assets	2671,047	3544,763	4346,594	2825,439
Client deposits	1921,287	2767,970	3336,875	2177,781
Loans	414,223	1550,413	2131,577	916,634

Table 11.

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		1665,679	2787,481	585,168
2	1665,679		1143,119	1126,578
3	2787,481	1143,119		2267,372
4	585,168	1126,578	2267,372	

Table 12.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Net profit	236,122	3	1678,767	3	,141	,929
Own funds	2856,043	3	2559,359	3	1,116	,465
Assets	843275,336	3	9500,245	3	88,764	,002
Client deposits	628462,814	3	41198,320	3	15,255	,025
Loans	966937,206	3	27875,836	3	34,687	,008

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 13.

Number of Cases in each Cluster

Cluster	1	3,000
	2	2,000
	3	1,000
	4	1,000
Valid		7,000
Missing		,000

From the presented data (Table 9) we see that now the first cluster is formed by years 2000, 2001 and 2002, the second by 2004 and 2005, the third by 2006 and the fourth only by the year 2003.

According to the data presented in the ANOVA table, assets once again have maximum influence in forming the clusters and net profit the least.

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

Supervisor: Snezhana Gocheva-Ilieva