Клъстерен анализ с SPSS: K-means анализ

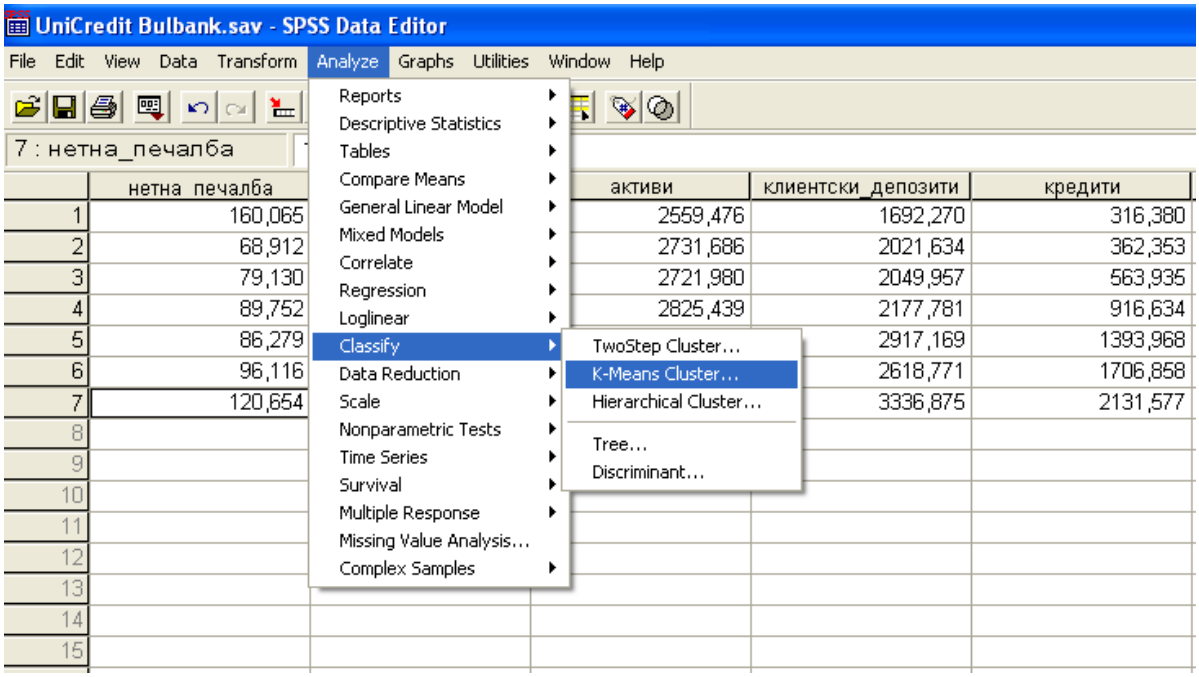
Клъстерният анализ е класификация без обучение, чиято цел е да се оформят естествени групи въз основа на много признаци едновременно. Целта при клъстерния анализ е n на брой обекта да се групират в k (k>1) на брой групи, наречени **клъстери**, като се използват p (p>0) на брой признаци (променливи). Самият клъстерен анализ е събирателно понятие и съдържа много на брой различни клъстеризационни процедури.

Едно важно деление на клъстеризационните процедури е в зависимост от това дали се задава предварително броят на клъстерите. При предварително зададен брой на клъстерите се използва метода **K-Means Cluster** (клъстерен анализ на K-средните). А когато броят на клъстерите не е предварително определен си служим с **Hierarchical Cluster Analysis** или т.н. йерархичен клъстерен анализ.

Голямото разнообразие на клъстеризационни процедури се поражда още от използваната метрика между различните обекти. По-известни метрики са: Евклидовото разстояние, Манхатъново разстояние, разстояние на Чебишев и др. Разнообразието се поражда и от използваните правила за създаване на клъстерите – за тяхното обединяване или разединяване.

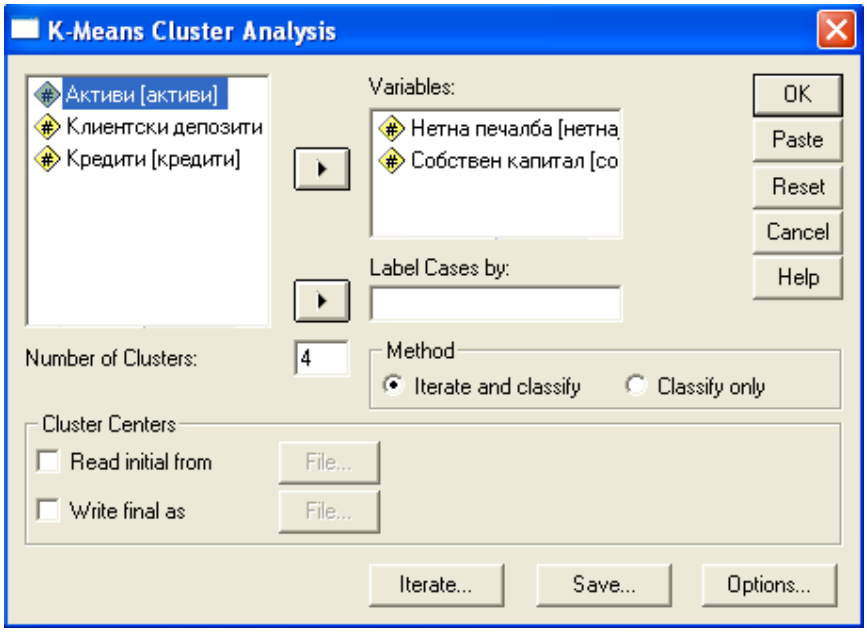
K-Means клъстерен анализ

От главното меню на SPSS се избира последователно **Analyze→ Classify → K-Means Cluster**.



фигура 1.

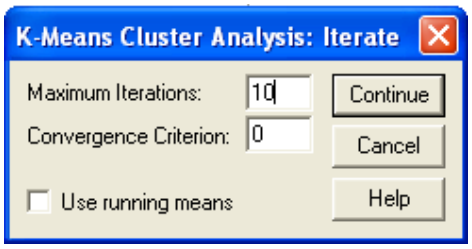
Маркират се променливите, въз основа на които ще се извърши клъстеризацията и се изпращат в полето за въвеждане на променливи величини **Variables**. А полето **Label Cases by** служи за въвеждане на стрингова променлива величина за маркиране на единиците. След това в полето **Number of Clusters** се указва броя на желаните клъстери. В нашия случай в полето **Method** се маркира **Iterate and Classify**. За разлика от другия метод – **Classify only**, който определя постоянни клъстерни центрове, този дава последователните итерации и на коя от тях се извършва финалната клъстеризация.



фигура 2.

В полето **Cluster Centers** се определя файла, който съдържа началните клъстерни центрове (ако има такъв) и файла, който да съдържа крайните клъстерни центрове (ако е необходимо). Където с **Read initial from** се определя файла, който съдържа началните клъстерни центрове, а **Write final as** определя файла, който да съдържа крайните клъстерни центрове.

С клавиша **Iterate** се определят критериите за актуализиране на клъстерните центрове, като с **Maximum Iterations** се указва максималният брой на итерациите (не повече от 999), а с **Convergence Criterion** критерият за сходство, който служи за прекратяване на итеративния процес. По подразбиране са дадени 10 итерации и критерий за сходство 0. Освен това тук е възможно да се маркира опцията **Use running means**. Ако тя бъде избрана, центровете на клъстерите се променят след присъединяването на всеки обект, а ако не бъде избрана центровете на клъстерите се изчисляват след като бъдат присъединени всички обекти към даден клъстер. В двата случая се получават различни резултати и затова трябва изрично да се указва как е осъществена клъстезацията. Продължава се с Continue.

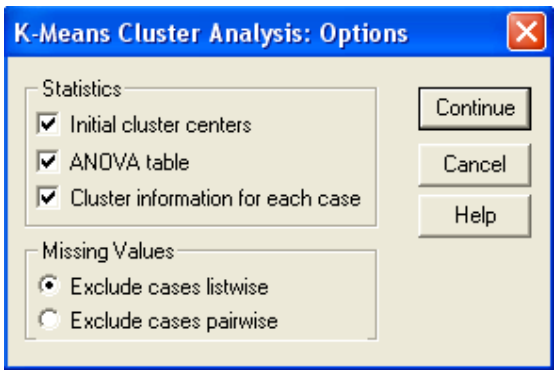


фигура 3.

С клавиша **Save** могат да се запишат във файла с данни нови променливи, показващи принадлежността на всеки обект към отделните клъстери (**Cluster Membership**) и разстоянието до центровете на клъстерите за всеки обект (**Distance from Cluster Center**).

фигура 4.

С клавиша **Options** се дава възможност за представяне на допълнителни статистики – началните клъстерни центрове (**Initial cluster centers**), таблица на дисперсионния анализ (**ANOVA table**) и информация за всеки обект за принадлежността му към даден клъстер (**Cluster information for each case**). Желателно е да се маркират и трите опции. Накрая с OK се получава резултата.



фигура 5.

Нека разгледаме накратко и без формули отделните етапи на клъстерния анализ с K-средни величини, използвайки данните от примера с UniCredit Bulbank, Таблица 1, (виж главата *Първи стъпки в SPSS*). Определяме броя на клъстерите на 4, като началните клъстерни центрове се оценяват от данните. За мярка на различие между единиците използваме квадратично евклидово разстояние. Като също така избираме центровете на клъстерите да се изчисляват след като бъдат присъединени всички обекти към даден клъстер, т.е. в полето **Use running means** не слагаме отметка.

Началните клъстерни центрове са представени в Таблица 1 (**Initial Cluster Centers**). Те представляват вектори със значенията по петте променливи величини, които се отнасят за 2000 година (първи клъстер), 2005 (втори клъстер), 2006 (трети клъстер) и 2003 (четвърти клъстер). Тези 4 години се намират на най-голямо разстояние по показатели една от друга.

Таблица 1.

Initial Cluster Centers

	Cluster			
	1	2	3	4
Нетна печалба	160,065	96,116	120,654	89,752
Собствен капитал	602,776	609,609	630,781	550,026
Активи	2559,476	3474,829	4346,594	2825,439
Клиентски депозити	1692,270	2618,771	3336,875	2177,781
Кредити	316,380	1706,858	2131,577	916,634

В Таблица 2 можем да видим броя на итерациите и промените в клъстерните центрове. При първата итерация 2001 г. се присъединява към 2000 г. и клъстерният център се актуализира. 2004 г. се присъединява към втория клъстер - 2005 година, а към четвъртия клъстер 2003 г. се присъединява 2002 г. Третият клъстер не се променя. При втората итерация процесът на преразпределение на единиците спира, понеже няма промени в клъстерните центрове.

Таблица 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.

В Таблица 3 са обобщени резултатите, т.е. коя единица към кой клъстер принадлежи, както и новите клъстерни центрове. Първият клъстер се формира от 2000 и 2001 година, вторият от 2004 и 2005, третият само от 2006 и четвъртият от 2002 и 2003 година.

В Таблица 4 можем да видим крайните клъстерни центрове, а разстоянието между крайните клъстерни центрове в Таблица 5.

Таблица 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

Таблица 4.

Final Cluster Centers

	Cluster			
	1	2	3	4
Нетна печалба	114,489	91,198	120,654	84,441
Собствен капитал	546,628	591,861	630,781	531,638
Активи	2645,581	3544,763	4346,594	2773,710
Клиентски депозити	1856,952	2767,970	3336,875	2113,869
Кредити	339,367	1550,413	2131,577	740,285

Таблица 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	

Ако сравним резултатите от Таблица 1 и Таблица 4 ще видим, че клъстерният център на третия клъстер не се променя. Това е така, защото той се състои само от един елемент – 2006 година.

Тъй като в нашия случай групите са формирани преднамерено в съответствие с разстоянието между тях в многомерното пространство, т.е. е нарушено условието за случайност на наблюденията в отделните групи, резултатите от дисперсионния анализ имат само описателен характер. С други думи не може да се използва равнището на значимост (колоната Sig. в табл. ANOVA – дисперсионен анализ на резултатите от клъстеризацията) за проверка на хипотезите относно средните величини. Въпреки това различията между F-отношенията (колоната F в табл. ANOVA) позволяват да се дадат най-общи заключения за ролята на отделните променливи величини при формиране на клъстерите.

В Таблица 6 са представени резултатите от дисперсионния анализ. Според тях най-голямо влияние при формирането на клъстерите имат активите, а най-малко – нетната печалба.

Таблица 6.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Нетна печалба	495,145	3	1419,744	3	,349	,795
Собствен капитал	2878,202	3	2537,200	3	1,134	,460
Активи	842788,443	3	9987,138	3	84,387	,002
Клиентски депозити	634017,636	3	35643,498	3	17,788	,021
Кредити	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.

Таблица 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

Таблица 7 представя данни както за броя на единиците във всеки клъстер, така и за общия брой и липсващите единици (ако има такива).

Сега ще представим резултатите от същата клъстеризационна процедура с тази разлика, че избираме центровете на клъстерите да се променят след присъединяването на всеки обект към даден клъстер и за тази цел маркираме опцията **Use running means**.

Таблица 8.

Iteration History(a)

	Change in Cluster Centers			
Iteration	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.

Таблица 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

Таблица 10.

Final Cluster Centers

	Cluster			
	1	2	3	4
Нетна печалба	102,702	91,198	120,654	89,752
Собствен капитал	535,501	591,861	630,781	550,026
Активи	2671,047	3544,763	4346,594	2825,439
Клиентски депозити	1921,287	2767,970	3336,875	2177,781
Кредити	414,223	1550,413	2131,577	916,634

Таблица 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	

Таблица 12.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Нетна печалба	236,122	3	1678,767	3	,141	,929
Собствен капитал	2856,043	3	2559,359	3	1,116	,465
Активи	843275,336	3	9500,245	3	88,764	,002
Клиентски депозити	628462,814	3	41198,320	3	15,255	,025
Кредити	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 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.

Таблица 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

От представените данни (Таблица 9) се вижда, че вече първият клъстер се формира от 2000, 2001 и 2002 година, вторият от 2004 и 2005, третият от 2006 и четвъртият единствено от 2003 година.

Според данните в Anova таблицата, отново най-голямо влияние при формирането на клъстерите оказват активите, а най-малко – нетната печалба.

Съставил: Десислава Войникова

Литература:

1. Манов, А. Б., Статистика със SPSS, изд. Тракия, София, 2001.
2. Гоев, В. Д. , Статистическа обработка и анализ на информацията от социологически, маркетингови и политически изследвания със SPSS, УНСС, София, 1996.