Klastrová analýza s SPSS: K-hodnotová klastrová analýza

Klastrová analýza je istým typom klasifikácie dát, ktorý sa používa separáciu dát do skupín. Cieľom klastrovej analýzy je kategorizácia n objektov do k (k>1) skupín nazývaných **klastre**, pomocou p (p>0) premenných. Klastrová (zhluková) analýza má niekoľko variant, tak ako mnoho iných štatistických analýz, pričom každá má svoje vlastné klastrovacie procedúry.

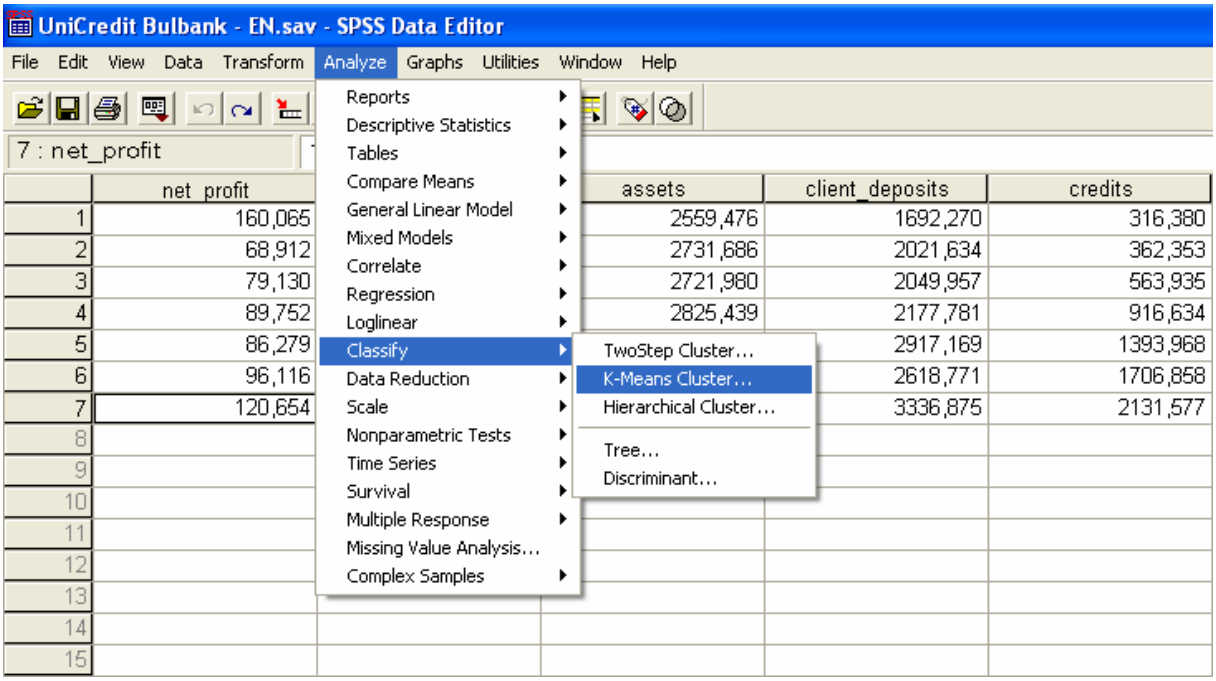
Klastrovacie procedúry sa dajú rozdeliť do dvoch základných podskupín. V prvom type procedúr sa počet klastrov preddefinuje. Táto metóda je známa ako **K-hodnotový klastering**. Ak počet klastrov nie je preddefinovaný, používame **Hierarchickú klastrovú** analýzu.

Veľká rozmanitosť klastrovacích procedúr je dôsledkom metrík, ktoré sú použité medzi jednotlivými objektmi. Najčastejšie používanými metrikami sú euklidovská metrika, Manhattanovská metrika, Čebyševova metrika, a iné. Pri tvorbe klastrov sa tiež používajú rôzne pravidlá. Niektoré povoľujú prvkom patriť do rôznych klastrov, pričom iné umožňujú byť prvkom iba jediného klastra.

K-hodnotová klastrová analýza

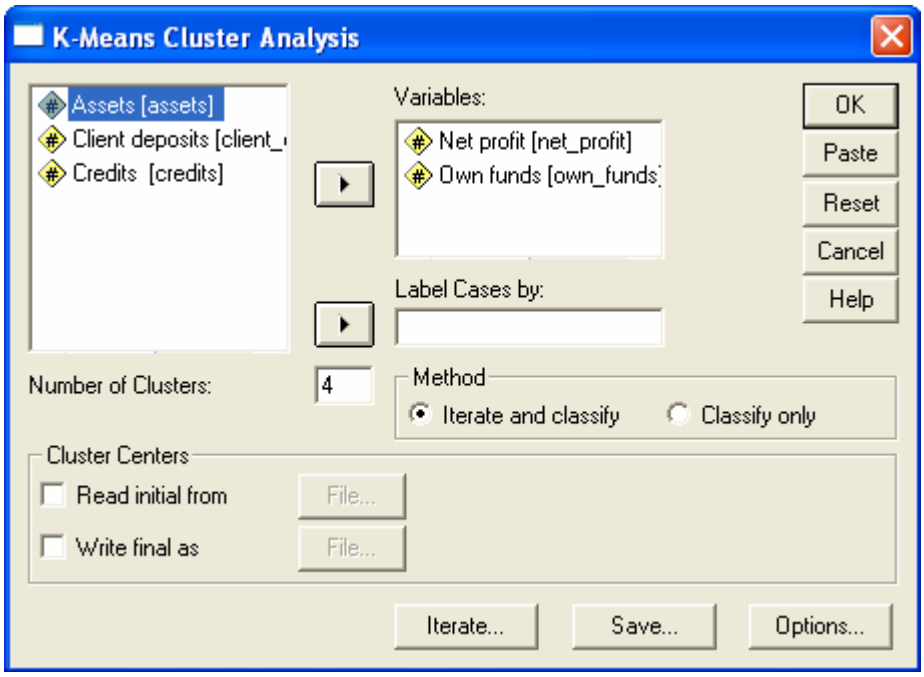
Z hlavného menu programu SPSS postupne vyberáme

Analyze→ Classify → K-Means Cluster.



Obrázok 1.

Označíme premenné, ktoré chceme roztriediť použitím klasteringu, a uložíme ich do **Variables –** okno na vkladanie premenných. Okno **Label Cases by** použijeme na vloženie reťazca opisujúceho jednotky. Potom určíme počet požadovaných klastrov v okne **Number of Clusters**. V našom prípade označíme v položke **Method** možnosť **Iterate and Classify**, ktorá na rozdiel od alternatívnej metódy – **Classify only** definujúcej pevné stredy klastrov, definuje postupné iterácie a určuje, ako sa bude vytvárať záverečné zhlukovanie do klastrov.

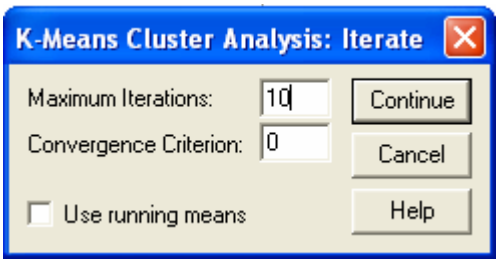


Obrázok 2.

V položke **Cluster Centers** špecifikujeme súbor (ak nejaký existuje), ktorý obsahuje začiatočné stredy klastrov a súbor (ak je potrebný), ktorý obsahuje konečné stredy klastrov. Voľbou **Read initial from** špecifikujeme súbor, ktorý obsahuje začiatočné stredy klastrov, a voľbou **Write final as** špecifikujeme súbor, ktorý obsahuje konečné stredy klastrov.

Tlačidlom **Iterate** môžeme určiť kritériá pre obnovovanie stredov klastrov, voľbou parametra **Maximum Iterations** zvolíme maximálny počet iterácií (nie viac ako 999), a pomocou **Convergence Criterion** rozhodneme, podľa akého pravidla sa má proces iterácie zastaviť. Nastavené hodnoty sú - 10 iterácií, kritérium konvergencie je 0. Môžeme tiež nastaviť podmienku **Use running means**. Znamená to, že stredy klastrov sa

menia po pridaní každého ďalšieho objektu. Ak nie je nastavená táto podmienka, stredy klastrov sa vypočítavajú po zaradení všetkých objektov do daných klastrov. V oboch prípadoch dostaneme iný výsledok, preto je potrebné špecifikovať spôsob, akým sa klastering dosiahne. Pokračujeme stlačením tlačidla Continue.

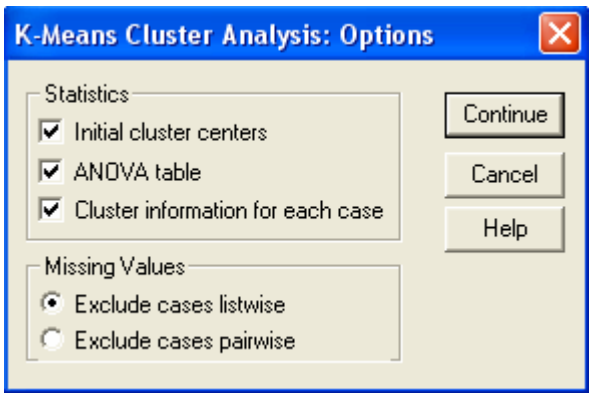


Obrázok 3.

Pomocou tlačidla **Save** môžeme do dátového súbor, ktorý určuje príslušnosť jednotlivých objektov ku klastrom, uložiť nové premenné (**Cluster Membership**), a aj vzdialenosť každého objektu od stredu klastra (**Distance from Cluster Center)**.

Obrázok 4.

V okne **Options** máme možnosť nastavovať ďalšie štatistické ukazovatele – začiatočné stredy klastrov (**Initial cluster centers**), tabuľka disperznej analýzy (**ANOVA table**) a informácia o príslušnosti každého objektu ku klastru (**Cluster information in each case?**). Spravidla nastavujeme všetky 3 hodnoty, výsledok získame po stlačení OK.



Obrázok 5.

Prejdeme stručne jednotlivými kroky K-hodnotovej klastrovej analýzy, použijeme dáta z príkladu o banke UniCredit Bulbank (Tabuľka 1 z kapitoly *Prvé kroky v SPSS*). Zvolíme počet klastrov 4 a nájdeme začiatočné stredy klastrov vyhodnotením dát. Použijeme štvorec euklidovskej vzdialenosti na určenie miery divergencie medzi jednotkami. Zvolíme tiež podmienku, aby sa stredy klastrov vypočítavali až po zaradení všetkých objektov do jednotlivých klastrov, t.j. neoznačíme **Use running means**.

Začiatočné stredy klastrov sú uvedené v Tabuľke 1 (**Začiatočné stredy klastrov**). Sú to vektory s hodnotami určenými piatimi premennými, ktoré sa vzťahujú na roky 2000 (prvý klaster), 2005 (druhý klaster), 2006 (tretí klaster) a 2003 (štvrtý klaster). Tieto 4 roky sú na maximálnom indexe vzájomnej vzdialenosti.

Tabuľka 1.

Začiatočné stredy klastrov

	Klaster			
	1	2	3	4
Čisté príjmy	160,065	96,116	120,654	89,752
Vlastné prostriedky	602,776	609,609	630,781	550,026
Aktíva	2559,476	3474,829	4346,594	2825,439
Vklady klientov	1692,270	2618,771	3336,875	2177,781
Úvery	316,380	1706,858	2131,577	916,634

V Tabuľke 2 vidíme počet iterácií a zmien stredov klastrov. V prvej iterácii sa k roku 2001 pridá rok 2000 a stred klastra sa nanovo prepočíta. Rok 2004 sa pridá do druhého klastra – rok 2005, a rok 2002 sa pridá do štvrtého klastra – rok 2003. Tretí klaster sa nezmení. V druhej iterácii sa proces redistribúcie zastaví, pretože nenastanú žiadne zmeny stredov klastrov.

Tabuľka 2.

Priebeh iterácie (a)

	Zmeny stredov klastrov			
Iterácia	1	2	3	4
1	200,730	227,959	,000	195,515
2	,000	,000	,000	,000

(a) Konvergence bola dosiahnutá vďaka žiadnym alebo minimálnym zmenám v stredoch klastrov. Maximálna absolútna súradnica kteréhokoľvek stredu je ,000. Aktívna iterácia je 2. Minimálna vzdialenosť medzi stredmi je 821,273.

Výsledky sú uvedené v Tabuľke 3, informácia, do ktorého klastra patrí každý objekt a nový stred klastrov. Prvý klaster tvoria roky 2000 a 2001, druhý obsahuje roky 2004 a 2005, tretí iba rok 2006 a štvrtý roky 2002 a 2003.

V Tabuľke 4 vidíme záverečné stredy klastrov, a v Tabuľke 5 – vzdialenosti výsledných stredov klastrov.

Tabuľka 3.

Príslušnosť ku klastrom

Číslo	Klaster	Vzdialenosť
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

Tabuľka 4.

Konečné stredy klastrov

	Klaster			
	1	2	3	4
Čisté príjmy	114,489	91,198	120,654	84,441
Vlastné prostriedky	546,628	591,861	630,781	531,638
Aktíva	2645,581	3544,763	4346,594	2773,710
Vklady klientov	1856,952	2767,970	3336,875	2113,869
Úvery	339,367	1550,413	2131,577	740,285

Tabuľka 5.

Vzdialenosti konečných stredov klastrov

Klaster	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	

Keď porovnáme výsledky z Tabuľky 1 a Tabuľky 4, všimneme si, že stred tretieho klastra sa nemení.

Pretože v našom prípade sa skupiny utvárajú ne základe vypočítavanej vzdialenosti objektov vo viacrozmernom priestore, akákoľvek náhodnosť pozorovaného javu v danej skupine neprichádza do úvahy, a výsledky disperznej analýzy sú iba opisné. Inými

slovami, nemôžeme použiť významnú hodnotu (Stĺpec Sign. v tabuľke ANOVA – disperzná analýza výsledkov klasteringu) na overenie hypotézy o strednej hodnote. Napriek tomu však rozdiely medzi F-priemermi (stĺpec F v tabuľke ANOVA) umožňujú načrtnúť všeobecné závery o význame rôznych stredných hodnôt pri formovaní klastrov.

V Tabuľke 6 sú uvedené výsledky disperznej analýzy. Ukazujú, že **aktíva** majú najväčší vplyv na utváranie klastrov a **Čisté príjmy** majú najmenší vplyv.

Tabuľka 6.

ANOVA

	Klaster		Chyba			
	Stredná kvadratická odchýlka	df	Stredná kvadratická odchýlka	df	F	Sig.
Čisté príjmy	495,145	3	1419,744	3	,349	,795
Vlastné prostriedky	2878,202	3	2537,200	3	1,134	,460
Aktíva	842788,443	3	9987,138	3	84,387	,002
Vklady klientov	634017,636	3	35643,498	3	17,788	,021
Úvery	957411,333	3	37401,709	3	25,598	,012

Výsledky F testov by mali slúžiť iba na opis situácie, pretože klastre boli zvolené tak, aby maximalizovali rozdiely medzi jednotlivými prípadmi v rôznych klastroch. Pozorované výsledky úrovní významnosti nie sú tomuto prispôsobené, a preto nemôžu byť interpretované ako testy hypotézy, že stredné hodnoty klastrov sú zhodné.

Tabuľka 7.

Počet prípadov v každom klastri

Klaster	1	2,000
	2	2,000
	3	1,000
	4	2,000
Platné údaje		7,000
Chýbajúce údaje		,000

Tabuľka 7 prezentuje dáta udávajúce počet jednotiek v každom klastri a ich výsledný počet a chýbajúce jednotky (ak nejaké exitujú).

Napokon uvedieme výsledky toho istého klastrovacieho procesu s tým rozdielom, že budeme požadovať prepočet stredov klastrov po pribudnutí každého nového objektu do niektorého klastra a vyberieme možnosť **Use running means**.

Tabuľka 8.

Priebeh iterácie(a)

Iterácia	Zmeny stredov klastrov			
	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) Proces iterácie sa zastavil, pretože sa vykonal maximálny počet iterácií. Proces iterácie nekonvergoval. Maximálna absolútna zmena súradnice pre ktorýkoľvek stred 0,005. Aktuálna iterácia je 10. Minimálna vzdialenosť medzi začiatočnými stredmi je 821,273.

Tabuľka 9.

Príslušnosť ku klastrom

Číslo		Klaster	Vzdialenosť
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

Tabuľka 10.

Konečné stredy klastrov

	Klaster			
	1	2	3	4
Čisté príjmy	102,702	91,198	120,654	89,752
Vlastné prostriedky	535,501	591,861	630,781	550,026
Aktíva	2671,047	3544,763	4346,594	2825,439
Vklady klientov	1921,287	2767,970	3336,875	2177,781
Úvery	414,223	1550,413	2131,577	916,634

Tabuľka 11.

Vzdialenosti medzi konečnými stredmi klastrov

Klaster	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	

Tabuľka 12.

ANOVA

	Klaster		Chyba			
	Stredná kvadratická odchýlka	df	Stredná kvadratická odchýlka	df	F	Sig.
Čisté príjmy	236,122	3	1678,767	3	,141	,929
Vlastné prostriedky	2856,043	3	2559,359	3	1,116	,465
Aktíva	843275,336	3	9500,245	3	88,764	,002
Vklady klientov	628462,814	3	41198,320	3	15,255	,025
Úvery	966937,206	3	27875,836	3	34,687	,008

Výsledky F testov by mali slúžiť iba na opis situácie, pretože klastre boli zvolené tak, aby maximalizovali rozdiely medzi jednotlivými prípadmi v rôznych klastroch. Pozorované výsledky úrovní významnosti nie sú tomuto prispôsobené, a preto nemôžu byť interpretované ako testy hypotézy, že stredné hodnoty klastrov sú zhodné.

Tabuľka 13.

Počet prípadov v každom klastri

Klaster	1	3,000
	2	2,000
	3	1,000
	4	1,000
Platné údaja		7,000
Chýbajúce údaja		,000

Z uvedených údajov (Tabuľka 9) sa dá zistiť, že tentokrát prvý klaster tvoria roky 2000, 2001 a 2002, druhý roky 2004 a 2005 a v treťom klastri je rok 2006, zatiaľ čo štvrtý klaster obsahuje iba rok 2003.

Podľa dát obsiahnutých v tabuľke ANOVA je zrejmé, že **aktíva** majú opäť najväčší vplyv pri formovaní klastrov a **Čistý zisk** zasa najmenší.