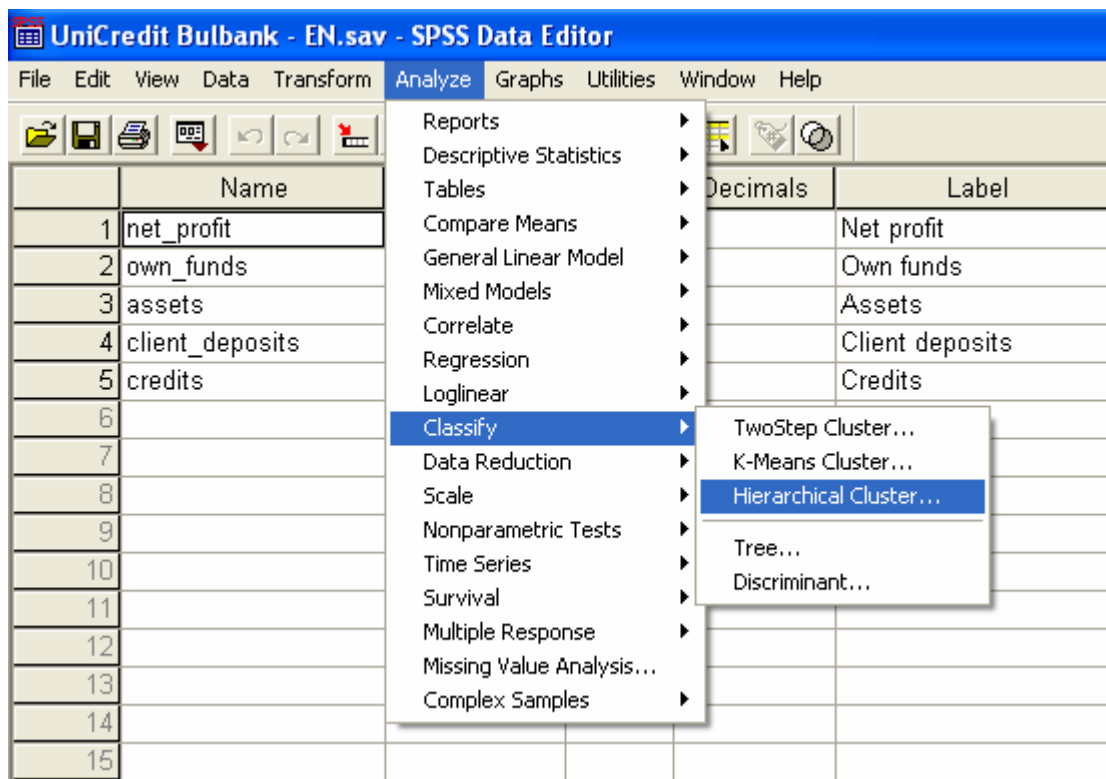


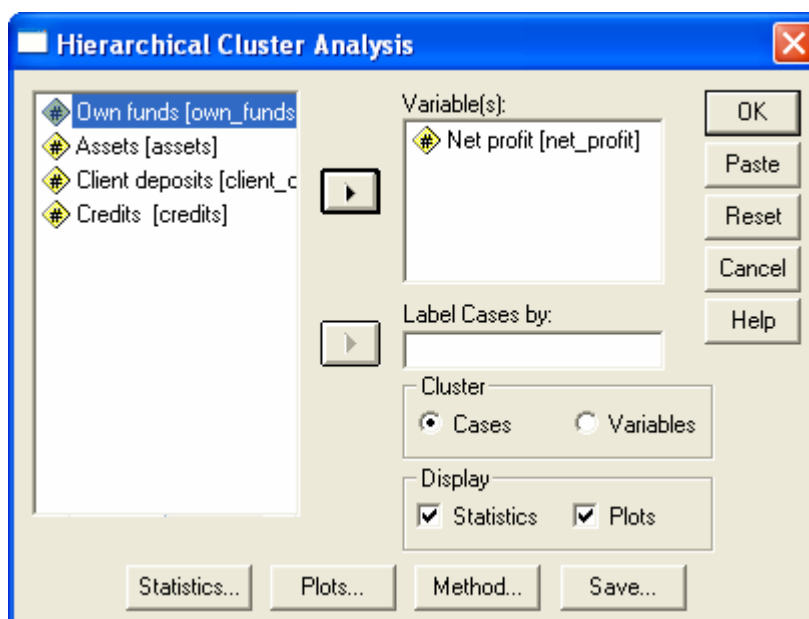
Klastrová analýza v SPSS: Hierarchická klastrová analýza

V hlavnom menu postupne volíme **Analyze** → **Classify** → **Hierarchical Cluster**.



Obrázok 1.

Objaví sa nasledujúce dialógové okno:

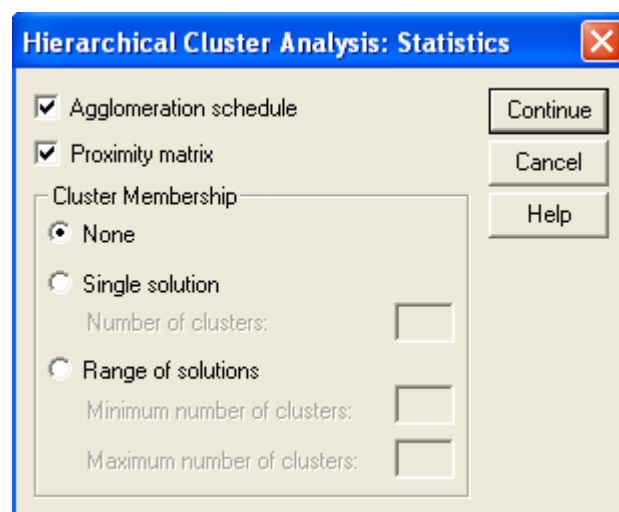


Obrázok 2.

Postupne po jednej vyberáme premenné, ktoré je potrebné analyzovať a ukladáme ich do okna **Variables**. Nasledujúce úkony vo veľkej miere ovplyvňuje typ analýzy,

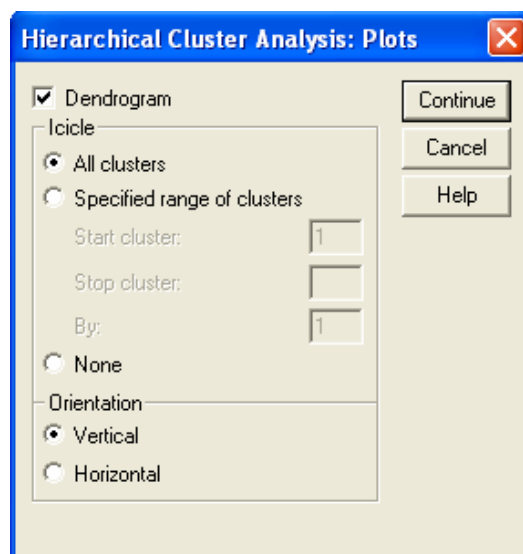
V okne **Cluster** môžeme vybrať jednu z možností - **Cases**, kde sa vykoná zhľukovanie objektov, a **Variables**, kde sa vykonáva zhľukovanie premenných. Ďalej je potrebné vybrať metódu pre identifikáciu objektov. Okno **Label Cases by** sa používa pre vkladanie reťazca, ktorý opisuje jednotky. Ak v okne **Cluster** nastavíme „Variables“, namiesto „Cases“ (objekty), potom musíme zadávať množinu premenných v zozname „Variables“ a okno „Label Cases“ nechať prázdne. Štandardne je v okne **Display** nastavené **Statistics**, čo umožňuje zobrazovanie výsledkov štatistickej analýzy, a **Plot**, na zobrazovanie grafov. V oboch prípadoch nie je potrebné odstrániť označenie.

V dolnej časti dialógového okna sa nachádzajú ďalšie štyri tlačidlá pre vkladanie doplňujúcich informácií. Aktivujeme tlačidlo **Statistics** (obrázok 2), ktoré sa používa na definovanie štatistických výsledkov zobrazovaných na obrazovke. Tu môžeme označiť **Agglomeration schedule** pre zobrazenie matice blízkosti, ktorá udáva informácie o vzdialenostiach medzi objektmi a klastrami. V prvom prípade sa zobrazuje postupnosť unifikácie – zjednocovania objektov v klastroch. Na začiatku považujeme každý objekt za samostatný klaster a až potom sa inicializuje vlastný proces zhľukovania do klastrov. Ďalej môžeme v okne **Cluster Membership** postupne voliť: **None** – ak nie je potrebné zobrazovať príslušnosť objektov ku klastrom; **Single solution** – ak je nastavený presný počet klastrov a **Range of solutions** – ak nastavíme rozsah požadovaných klastrov – minimálny a maximálny počet klastrov, ktoré chceme získať. Pokračujeme stlačením tlačidla **Continue**.



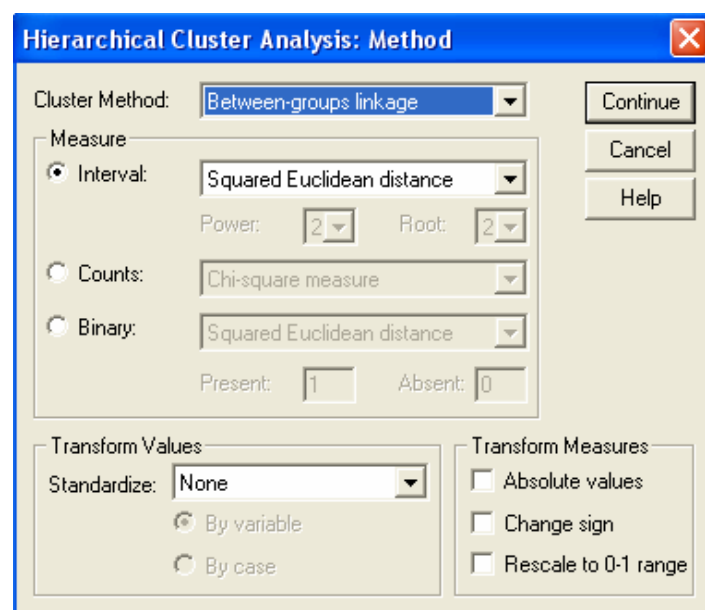
Obrázok 3.

Pri aktivácii tlačidla **Plots** môžeme zvoliť ponuku **Dendrogram**, ak chceme získať aj grafickú vizualizáciu výsledkov hierarchickej klastrovej analýzy. Dendrogram je stromový graf, v ktorom každý uzol reprezentuje jednu fázu zhukovacieho procesu. Dáva ďalšiu informáciu o významnosti vzdialenosti medzi dvoma klastrami v momente ich zjednocovania. Horizontálna čiarkovaná čiara dendrogramu indikuje vzdialenosť v upravenej mierke, v ktorej sú klastre vytvárané. V okne **Icicle** zaškrtnutím **All clusters** nastavíme podmienku zabezpečujúcu, aby diagram zahŕňal všetky klastre, použitím **Specified range of clusters** môžeme špecifikovať počet klastrov a pomocou **None** zrušíme „Icicle“. Pomocou tlačidiel **Orientations** môžeme zvoliť typ diagramu – horizontálny alebo vertikálny.



Obrázok 4.

Po aktivácii **Method** sa v okne **Hierarchical Cluster Analysis** objaví nasledujúce dialógové okno:



Obrázok 5.

Najprv špecifikujeme pomocou **Cluster Method** klastrovaciu metódu, ktorú mienime použiť. V programe SPSS máme k dispozícii 7 metód:

- Metóda spájania medzi skupinami
- Metóda spájania v rámci skupín
- Metóda najbližšieho suseda
- Metóda najvzdialenejšieho suseda
- Metóda stredového zhukovania
- Metóda zhukovania podľa strednej hodnoty
- Wardova metóda

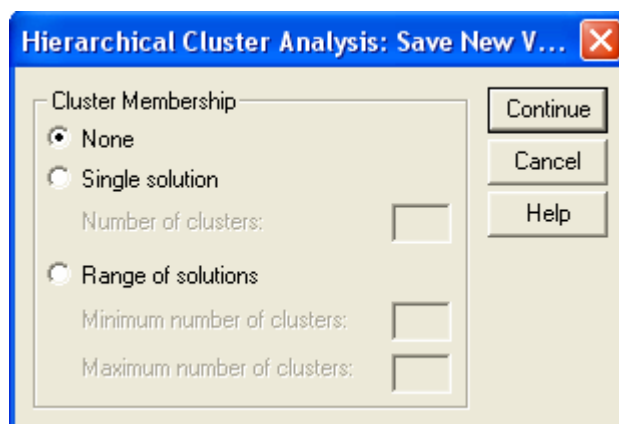
Každá z uvedených metód vedie k inému výsledku klasteringu. Nedá sa povedať, ktorá z nich je najlepšia, ale ak hľadáme klastre v tvare „reťazca“, je vhodné použiť metódu spájania medzi skupinami a metódu najbližšieho suseda. Keď hľadáme klastre v tvare „klastra“, odporúča sa použiť metódu spájania v rámci skupín a metódu najvzdialenejšieho suseda. Je treba pripomenúť, že na rozdiel od klastra v tvare „klastra“ v prípade klastra v tvare „reťazca“ sa počet objektov v jednotlivých klastroch výrazne líši.

V okne **Measure** musíme určiť podmienku konverencie, t.j. metódu merania podobnosti a odlišnosti medzi jednotkami. Zvolíme ju podľa mierky a použitých premenných – či sa jedná o interval (**Interval**), kategóriu (**Counts**) alebo číslo (**Binary**). Inými slovami, určíme mieru podobnosti a odlišnosti číselných, nečíselných alebo alternatívnych premenných. V okne **Transform Values** pomocou **Standardize** určujeme metódu šandardizácie premenných. Tieto metódy sme opísali v predchádzajúcom texte.

V okne **Transform Measures** môžeme použitím **Absolute Values** eliminovať smer previazanosti, ak používame korelačný koeficient pri zoskupovaní premenných. Pomocou **Change Sign** zmeníme znamienko pred mierou podobnosti a tým ju zmeníme na mieru odlišnosti - divergencie. Použitím **Rescale to 0-1 Range** šandardizujeme mieru podobnosti a odlišnosti na interval od 0 do 1, ak je to potrebné.

Použitím tlačidla **Save** uložíme výsledky klastrovej analýzy, t.j. príslušnosť jednotlivých objektov ku odpovedajúcim klastrom, ako ďalšiu premennú dátového súboru. Tu môžeme určiť metódu ukladania informácie - opisu príslušnosti ku klastrom pre jednotlivé objekty. **None** zruší ukladanie týchto opisov pre jednotlivé objekty.

Pomocou **Single Solution** môžeme uložiť opis príslušnosti ku klastru pre každú jednotku, ak je počet klastrov určený vopred, a pomocou **Range of Solutions** uložíme opis príslušnosti ku klastru tak, že využijeme preddefinované postupnosti riešení zhľukovania.



Obrázok 6.

Vráťme sa späť k príkladu so základnými ukazovateľmi banky UniCredit Bulbank a vykonajme hierarchickú klastrovú analýzu použitím priemerovaného spájania medzi skupinami (Average Linkage Between Groups). V Tabuľke 1 je zobrazený výsledok, t. j. platné a chýbajúce údaje a celkové hodnoty.

Tabuľka 1.

Súhrn spracovaných výsledkov (a)

Prípady					
Platné údaje		Chýbajúce údaje		Celkové hodnoty	
N	Percent	N	Percent	N	Percent
7	100,0	0	,0	7	100,0

(a) Average Linkage (Between Groups)

V Tabuľke 2 je uvedená matica odlišností (the Proximity Matrix), ktorú vytvoríme priamo v programe SPSS. Matica obsahuje štvorce euklidovských vzdialeností s mierou divergencie zodpovedajúcej údajom z príkladu. Štvorec vzdialenosti medzi dvoma prvými rokmi sa vypočíta podľa vzťahu:

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

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

Tabuľka 2.

Proximity Matrix

Prípád	Štvorce euklidovských vzdialeností						
	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

Toto je matica odlišnosti

Euklidovská vzdialenosť do značnej miery závisí od miery a rozsahu rôznych premenných. Premenná, ktorá je vyjadrená vyššími číslami, viac ovplyvní jej výpočet.

Pozrime sa pozornejšie na proces hierarchického zhľukovania pomocou metódy priemerovaného spájania medzi skupinami, potom, ako sme získali maticu vzdialeností.

	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	

V prvom kroku spájania sa skombinuje druhý a tretí rok, pretože vzdialenosť medzi nimi je aspoň $s_{23} = 42\,154,582$. Rozmery matice vzdialeností sa zredukujú o 1 a matica má nasledujúce prvky:

	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	

V ďalšom kroku sa skombinujú - prvý a druhý rok (2000 a 2001, 2002), pretože podľa teórie získame najmenšiu strednú vzdialenosť: $s_{12} = 391\,366,507/2 = 195\,683,253$.

Rozmery matice vzdialeností sa opäť zredukujú o 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	

V treťom kroku sa skombinujú - tretí a štvrtý klaster (2004 and 2005),

pričom: $s_{34} = 207\,861,100$.

	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	

V ďalšom kroku skombinujeme prvý a druhý klaster (2000, 2001, 2002 a rok 2003). V tomto prípade $s_{12} = 1\,171\,772,380/3 = 390\,590,794$.

	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	

V piatom kroku skombinujeme druhý a tretí klaster, teda 2004, 2005 a 2006, pričom stredná vzdialenosť je aspoň $s_{23} = 2\,717\,372,028/2 = 1\,358\,686,014$.

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

V poslednom kroku skombinujeme dva zostávajúce klastre, pričom stredná vzdialenosť je $s_{12} = 48\,485\,654,420/12 = 4\,040\,471,202$.

Výsledky jednotlivých krokov hierarchickej klastrovej analýzy v SPSS zosumarizujeme a zobrazíme v tabuľke nazývanej **Agglomeration Schedule** – **aglomeračná schéma**. V tejto tabuľke je tiež stĺpec s doteraz vypočítanými strednými vzdialenosťami. V tomto prípade je štvorec euklidovskej vzdialenosti použitý ako miera.

Tabuľka 3.

Agglomeration Schedule

Krok	Skombinované klastre		Koeficienty	Prvý výskyt klastra		Ďalší krok
	Klaster 1	Klaster 2		Kluster 1	Kluster 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

Prvý stĺpec **Krok** aglomeračnej schémy zobrazuje počet jednotlivých krokov a v poslednom kroku sú všetky analyzované objekty skombinované do jedného klastra. Vo väčšine prípadov je ich $n-1$. Stĺpce so spoločným názvom **Skombinované Klastre** ukazuje počty klastrov, ktoré sú skombinované v rôznych krokoch. Napríklad v prvom kroku je skombinovaný druhý a tretí klaster. Stĺpec **Coefficients** udáva stredné vzdialenosti (viď. teória) pre kombinovanie klastrov. Tieto koeficienty závisia od zvolenej metódy pre tvorbu klastrov. Hodnoty v tomto stĺpci sa môžu použiť pre približné vyhodnotenie podobnosti klastrov vytvorených v jednotlivých krokoch. Veľké koeficienty (pre miery divergencie) alebo malé koeficienty (pre miery podobnosti) indikujú, že klaster je relatívne heterogénny a obsahuje jednotky, ktoré sa od seba výrazne líšia. Koeficienty v tomto stĺpci tiež slúžia na získanie približného prehľadu o počte klastrov, ktoré je treba vyprofilovať z praktického hľadiska. Preto vyšetrujeme tiež také kroky, v ktorých sa objaví náhla výrazná zmena koeficientov. Stĺpce so spoločným názvom **Prvý výskyt klastra** zobrazujú kroky, v ktorých sa daný klaster objavuje po prvý raz a stĺpec **Ďalší krok** indikuje krok, v ktorom sa daný klaster objaví znovu a skombinuje sa s iným klastrom. V prvom kroku sa napríklad skombinovali druhý a tretí klaster, vytvorí sa nový klaster a je mu priradené číslo 2. V druhom kroku je skombinovaný novo vytvorený klaster 2 s klastrom 1, atď.

Výsledky jednotlivých krokov hierarchickej klastrovej analýzy sa dajú tiež ilustrovať pomocou tkzv. **Icicle Plot (cencúľový diagram)**. Môže byť vertikálny alebo horizontálny. Obrázok 7 prezentuje vertikálny diagram výsledkov klastrovej analýzy rozvoja a výsledkov banky UniCredit Bulbank za jednotlivé roky.

Vertikálny cencúľový diagram

Počet klastrov	Prípad												
	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

Obrázok 7.

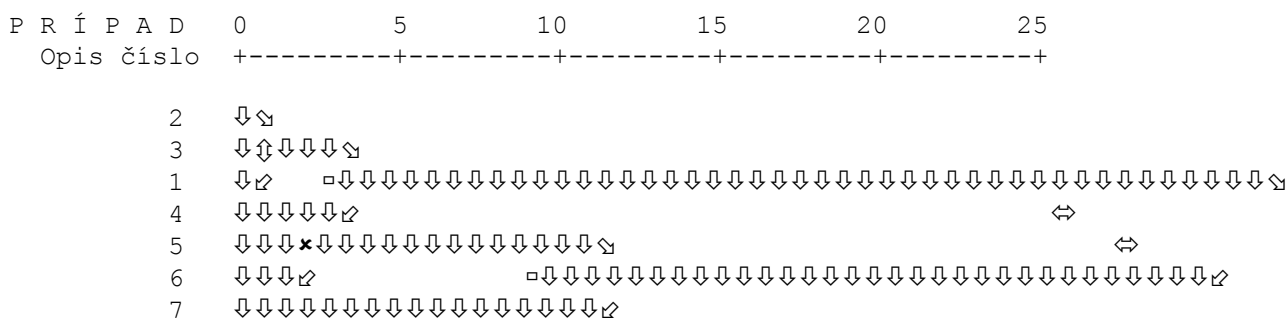
Riadky diagramu zodpovedajú počtu klastrov. Vo väčšine prípadov ich je $n-1$. Samostatný stĺpec, ktorý obsahuje X vo všetkých riadkoch zodpovedá jednotlivým jednotkám. Medzi stĺpcami, ktoré zodpovedajú jednotkám, sú ďalšie stĺpce, označené inak. Cencúľ je viditeľný zdola nahor. Napríklad roky 2001 a 2002 sú skombinované v klastrí 6 – stĺpec medzi nimi je označený až po posledný riadok, t.j. je v ňom plný počet, 6 znakov X. Rok 2000 je pridaný k roku 2001 a 2002, a tvoria klaster 5, preto stĺpec medzi 2000 a 2001 obsahuje päť znakov X, atď. Keď je počet kombinácií relatívne veľký, informácie vo vertikálnom cencúľovom diagrame sa ukazujú ako nedostatočné. V tomto prípade je potrebné použiť horizontálny cencúľový diagram, alebo graficky interpretovať časti klastrov.

Na grafickú interpretáciu výsledkov hierarchickej klastrovej analýzy môžeme použiť spomínaný dendrogram.

* * * * * H I E R A R C H I C K Á K L A S T R O V Á A N A L Ý Z A * * * * *

Dendrogram používajúci priemerované spájanie (medzi skupinami)

Prepočítaná Vzdialenosť skombinovaných Klastrov



Obrázok 8.

Horizontálna bodkovaná čiara dendrogramu na obrázku 8 reprezentuje vzdialenosti prepočítané v mierke, v ktorej sa klaster kombinujú. Minimálna vzdialenosť – v tomto

prípade 42 154,582 zodpovedá hodnote 1, a maximum – 4 040 471,202 zodpovedá hodnote 25. Dendrogram oprávňuje sformulovať nasledujúce výsledky:

 Roky 2000, 2001 a 2002 sú skombinované do spoločného klastra s relatívne malou vzdialenosťou, t.j. klaster je relatívne homogénny;

 Rok 2003 tvorí samostatný klaster, ktorý je postupne skombinovaný s klastrom 2000, 2001 a 2002 s relatívne rovnakými indexmi a takmer dvojnásobnými úvermi.

 Roky 2004 a 2005 tvoria samostatný klaster a sú skombinované s klastrom 2006, a sú vo výrazne väčšej vzdialenosti v porovnaní s ostatnými. Vykazujú tiež relatívne vyššie hodnoty všetkých základných ukazovateľov, najmä rok 2006.

Všimnime si, že výsledok získaný hierarchickou klastrovou analýzou je rovnaký, ako výsledok, ktorý sme získali K-hodnotovou klastrovou analýzou pri zmene stredov klastrov po pripojení každého objektu k danému klastru.