Postup analýzy shluků

Poskytuje empirické a objektivní metody
ke klasifikaci objektů

1. krok: Cíle analýzy shluků
2. krok: Formulace úlohy analýzy shluků
3. krok: Předpoklady analýzy shluků
4. krok: Výstavba dendrogramu shluků
5. krok: Interpretace shluků
6. krok: Validace a profilování shluků

1. krok: Cíle analýzy shluků

Rozdělení objektů do shluků dle podobnosti objektů
a dle specifikovaných vlastností - proměnných.

Popis systematiky (taxonomie): empirická klasifikace.
Shluky objektů jsou porovnány s jejich teoretickou typologií.

Zjednodušení dat: zjednodušený pohled na soubor objektů.
Na oddělené shluky objektů se hledí dle jejich vlastností.

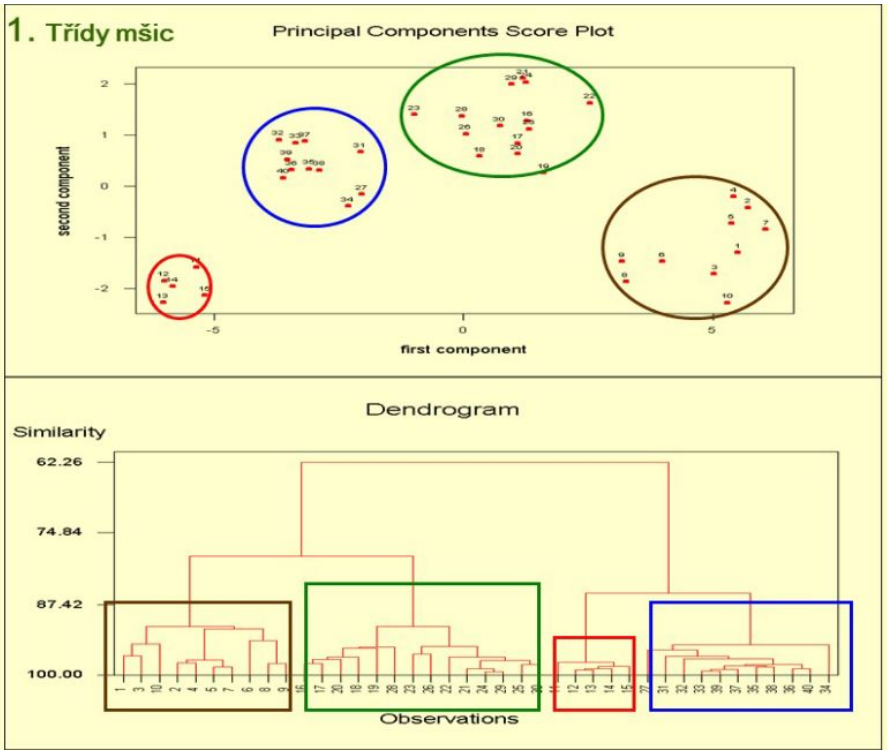
Identifikace vztahu: dle struktury shluků je snadnější odhalit vztahy mezi objekty. Shluky mohou být předmětem dalšího kvalitativního uvažování.

Úloha 1. Klasifikace polétavých mšic (Kompendium B404)

Jeffers (1967)25 studoval 40 jedinců polétavých mšic *(Alate adelges)* : 19 ukazatelů k rozlišení druhů, 14 znaků délky a šířky, 4 znaky se týkají počtu a 1 binární vyjadřuje přítomnost či absenci: x1 délka těla, x2 šířka těla, x3 délka předního křídla, x4 délka zadního křídla, x5 počet průduchů, x6 délka tykadla I, x7 délka tykadla II, x8 délka tykadla III, x9 délka tykadla IV, x10 délka tykadla V, x11 počet tykadlových ostnů, x12 délka posledního článku nohy, x13 délka holeně, tibia, x14 délka stehna, x15 délka sosáku, x16 délka kladélka, x17 počet kladélkových trnů, x18 řitní otvor, x19 počet háčků zadních křídel

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19
21.2	11	7.5	4.8	5	2	2	2.8	2.8	3.3	3	4.4	4.5	3.6	7	4	8	0	3
20.2	10	7.5	5	5	2.3	2.1	3	3	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10	7	4.6	5	1.9	2.1	3	2.5	3.3	1	4.2	4.4	3.3	7	4	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
20.6	11	8	4.8	5	2.4	2	2.9	2.7	3	4	4.2	4.7	3.5	6.7	4	6	0	3
19.1	9.2	7	4.5	5	1.8	1.9	2.8	3	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4	8	0	3
15.5	8.2	6.3	4.9	5	2	2	2.9	2.4	3	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
19.7	9.9	8.2	4.7	5	2.2	2	3	3	3.1	0	4.1	4.3	3.3	6	3.8	8	0	3
10.6	5.2	3.9	2.3	4	1.2	1	2	2	2.2	6	2.5	2.5	2	4.5	2.7	4	1	2
9.2	4.5	3.7	2.2	4	1.3	1.2	2	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
9.6	4.5	3.6	2.3	4	1.3	1	1.9	1.7	2.2	4	2.4	2.3	1.7	4	2.3	4	1	2
8.5	4	3.8	2.2	4	1.3	1.1	1.9	2	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
11	4.7	4.2	2.3	4	1.2	1	1.9	2	2.2	4	2.5	2.5	2	4.5	2.6	4	1	2
18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6	4.5	9	1	2
17.6	8.3	6	3.8	5	2	1.9	2	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2
19.2	8.6	6.2	3.4	5	2	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
15.4	7.6	7.1	3.4	5	2	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6	4.2	8	1	3
15.1	7.3	6.2	3.8	5	2	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6	4.5	0	1	2
19.1	8.8	6.4	3.9	5	2.2	2	2.3	2.4	2.9	4	3.8	4	3	6.5	4.5	0	1	2.5
15.3	6.4	5.3	3.3	5	1.7	1.6	2	2.2	2.5	5	3.4	3.4	2.6	5.4	4	0	1	2
14.8	8.1	6.2	3.7	5	2.2	2	2.2	2.4	3.2	5	3.5	3.7	2.7	6	4.1	0	1	2
16.2	7.7	6.9	3.7	5	2	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	0	1	2.5
13.4	6.9	5.7	3.4	5	2	1.8	2.8	2	2.6	4	3.6	3.6	2.6	5.5	3.9	0	1	2
12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3	2.2	5.1	3.6	9	1	3
12	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3	5	3.3	3.5	2.6	5.4	4.3	8	1	2
14.1	7	5.5	3.6	5	2.2	2	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	0	1	2
16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6	4	0	1	2.5
14.1	5.4	5	3	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
10	6	4.2	2.5	5	1.6	1.4	1.4	2	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2	5.1	3.7	8	0	2
13	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5	3.6	8	1	2
12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2	5	3.2	6	1	2
12	5.4	4.9	3	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2	4.2	3.7	6	1	2
10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2	5	3.5	8	1	2
11.1	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5	3.1	8	1	2

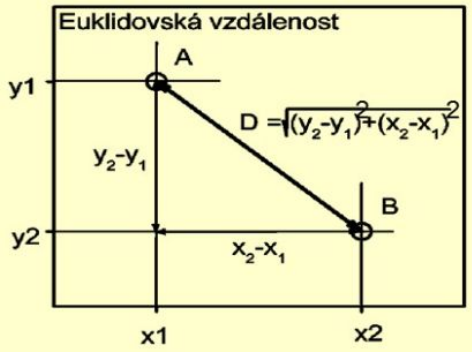
Úloha 1. Třídy mšic

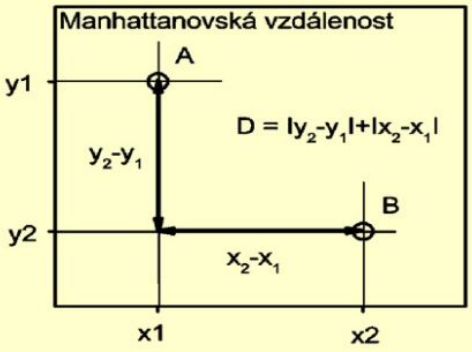


2.3 Míry podobnosti

Podobnost je měřena rozličnými způsoby

Míry vzdálenosti: nejčastěji užívané míry podobnosti. Vzdálenost je reciproká hodnota podobnosti. Čím větší hodnota vzdálenosti, tím menší podobnost.





Eukleidovská vzdálenost zvaná také *geometrická metrika*

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

Manhattanská vzdálenost zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika* je definovaná

$$d_H(x_k, x_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}|$$

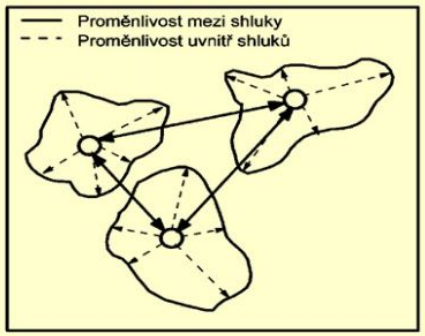
4. krok: Výstavba dendrogramu shluků

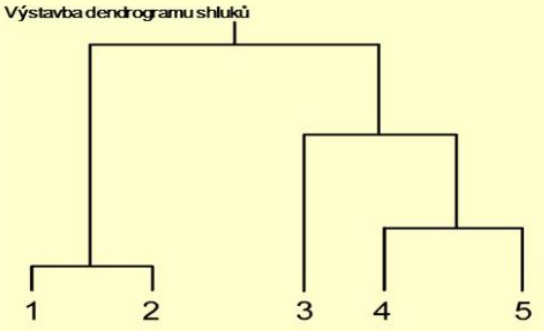
Vedle algoritmu je třeba vybrat i vhodný postup.

Rozlišovací kritérium: maximalizace rozdílů mezi shluky, Proměnlivost mezi shluky vůči proměnlivosti uvnitř shluků.

Test: poměr roztylu mezi shluky vůči průměru rozptylu uvnitř shluků

Algoritmy se dělí: hierarchické a nehierarchické.





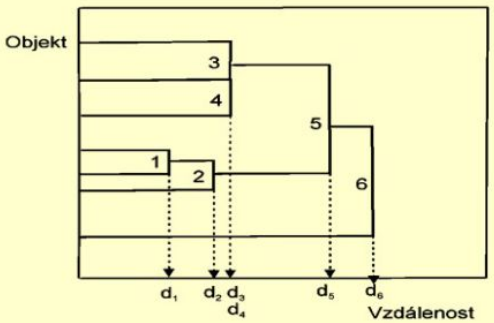
Hierarchické shlukování

konstrukce stromovité struktury, dendrogramu

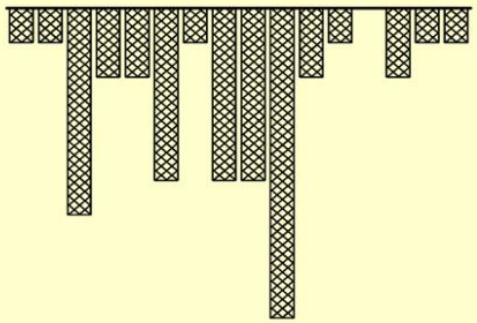
Způsoby hierarchického shlukování: aglomerační a divizní,

Aglomerační způsob: nejprve se spojí dva nejbližší objekty v jediný shluk, pak se připojí třetí objekt k prvním dvěma objektům a vznikne společný shluk. Tak se seskupí všechny objekty do jednoho velikého shluku.

1) růstový strom (dendrogram),



2) vertikální krápníkovitý diagram,



Aglomerační způsoby (algoritmy) výstavby dendrogramu shluků:

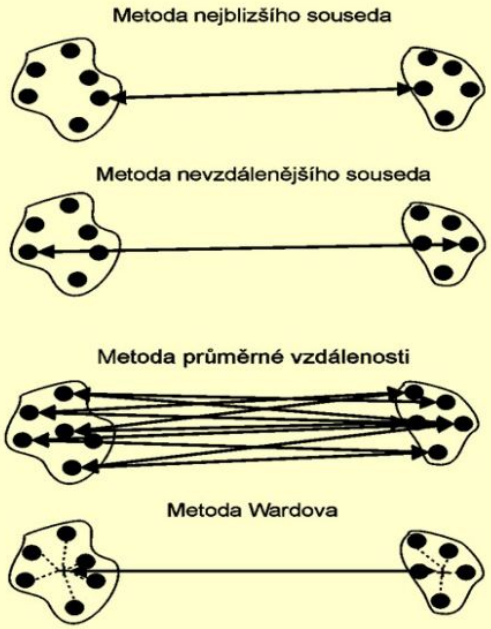
Metoda nejbližšího souseda: je postavena na minimální vzdálenosti objektů.

Metoda nejvzdálenějšího souseda: je postavena nikoliv na minimální ale na maximální vzdálenosti.

Metoda průměrového linkování: kritériem je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

Wardova metoda: vzdálenost mezi dvěma shluky je tvořena na základě sumy čtverců přes všechny proměnné mezi dvěma shluky.

Metoda těžiště: vzdálenost těžišť shluků spojených Euklidovskou vzdáleností nebo čtvercem Euklidovské vzdálenosti. Těžiště shluku je průměrná hodnota objektů v proměnných, vyjádřená ve shlukových proměnných.



Míra věrohodnosti“nejlepšího dendrogramu”

1. kritérium těsnost proložení:

kofenetický korelační koeficient CC

- nejlépe odpovídá struktuře objektů a znaků mezi objekty,
- je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

2. kritérium těsnosti proložení:

kritérium delta Δ

- měří stupeň přetvoření struktury dat,
- je žádoucí, aby hodnoty *delta* byly blízké nule,
- je definováno

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} | d_{jk} - d^{*}_{jk} |^{A}}{\sum_{j<k}^{N} (d^{*}_{jk})^{A}} \right]^{1/A}$$

kde A = 0.5 nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^{*}_{ij} je vzdálenost získaná z dendrogramu.

Úloha 2. Vytvoření dendrogramu objektů neuroleptika (Kompendium B402)

Liší se v účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění. Provedeme klasifikaci neuroleptik do shluků podobných účinků s ohledem na 4 znaky.

Data: Charakter proměnných (převrácená hodnota mediánové účinné dávky 1/ED50 [kg/mg]): **B402x1** značí název neuroleptika, **B402x2** je pro potlačení nervozity, **B402x3** značí potlačení stereotypního chování, **B402x4** je pro potlačení záchvatu a třesu, a **B402x5** znamená dávku smrtícího účinku.

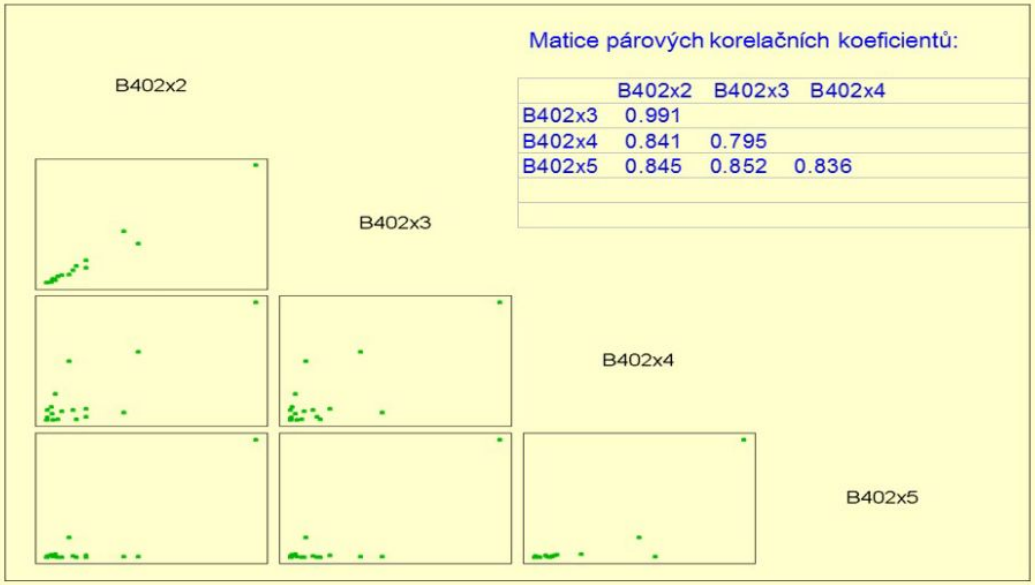
B402x1	B402x2	B402x3	B402x4	B402x5
1 Chlorphromazine	3.846	3.333	1.111	1.923
2 Promazine	0.323	0.213	0.108	1.429
3 Trifluperazine	27.027	17.857	0.562	0.14
4 Fluphenazine	17.857	15.385	1.695	1.075
5 Perphenazine	27.027	27.027	1.961	2.083
6 Thioridazine	0.244	0.185	0.093	1.333
7 Pifluthixol	142.857	142.857	20.408	163.934
8 Thiothixene	4.348	4.348	0.047	0.345
9 Chorprothixene	5.882	2.941	4.545	4.167
10 Spiperone	62.5	47.619	11.765	0.847
11 Haloperidol	52.632	62.5	1.282	0.568
12 Azaperone	2.941	1.282	2.222	3.03
13 Pipamperone	0.327	0.187	1.724	0.397
14 Pimozide	20.408	20.408	0.107	0.025
15 Metitepine	15.385	10.204	10.204	27.027
16 Clozapine	0.161	0.093	0.327	0.323
17 Perlapine	0.323	0.323	0.37	0.067
18 Sulpiride	0.047	0.047	0.003	0.001
19 Butaclamol	10.204	9.091	1.471	0.025
20 Molindone	7.692	7.692	0.14	0.006

Diagram korelační matice

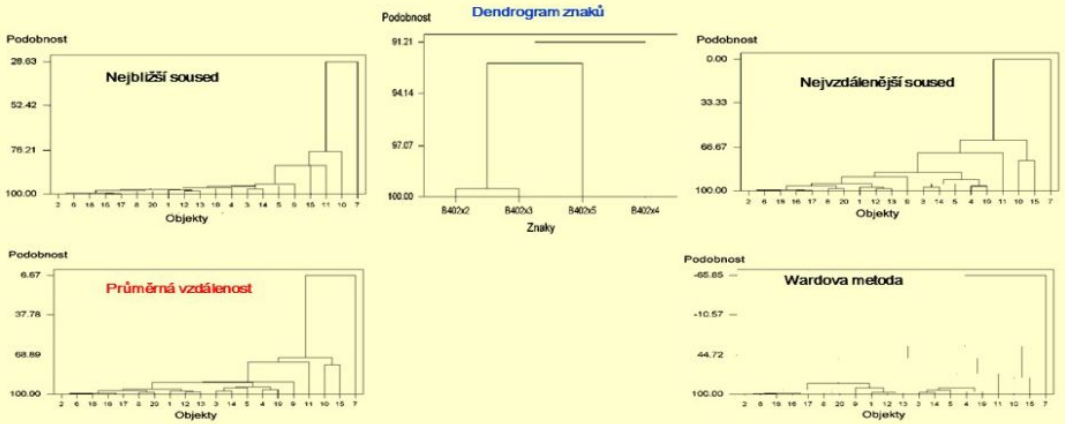
a statistická významnost korelace pomocí Pearsonových párových korelačních koeficientů

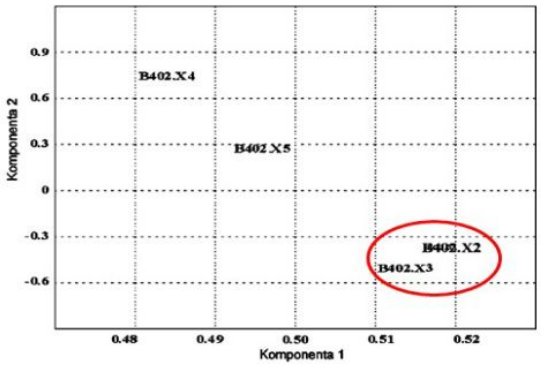
Matice párových korelačních koeficientů:

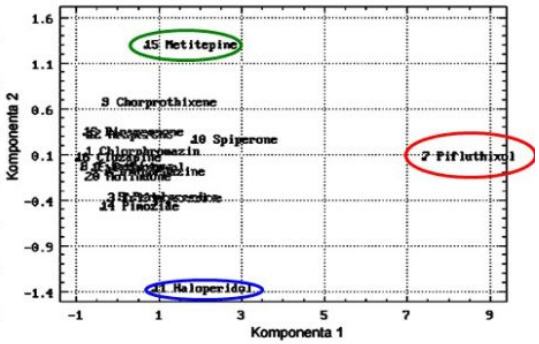
	B402x2	B402x3	B402x4
B402x3	0.991		
B402x4	0.841	0.795	
B402x5	0.845	0.852	0.836

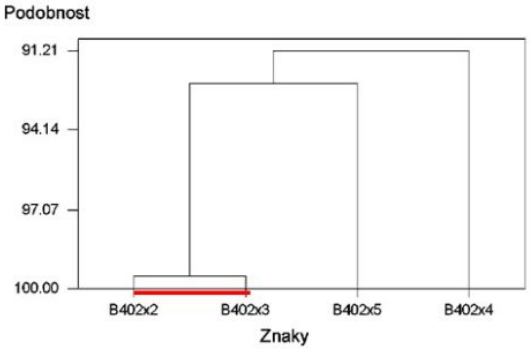


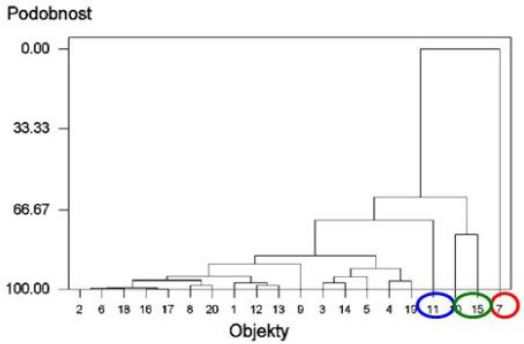
1. Metoda shlukování: **Skupinový průměr,** *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.987356, *Delta(0.5):*0.137455, *Delta(1.0):*0.125290;
2. Metoda shlukování: Jednoduchý průměr, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:*0.988876, *Delta(0.5):*0.177810, *Delta(1.0):*0.188781;
3. Metoda shlukování: Těžiště, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:*0.984750, *Delta(0.5):*0.175238, *Delta(1.0):*0.166599;
4. Metoda shlukování: Nejbližšího souseda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:*0.988598, *Delta(0.5):*0.474238, *Delta(1.0):*0.391993;
5. Metoda shlukování: Median, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:*0.984215, *Delta(0.5):*0.452308, *Delta(1.0):*0.428346;
6. Metoda shlukování: Wardova metoda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.979285, *Delta(0.5):*0.549394, *Delta(1.0):*0.492716.







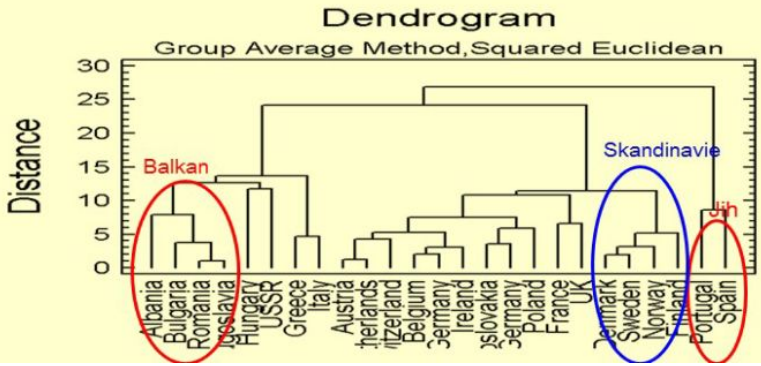




5. krok: Interpretace shluků

(a) Vyšetření každého shluku v pojmech shlukových proměnných.
(b) Pojmenování shluků nebo jeho označení, které vystihuje jeho podstatu a povahu.

Profilování a interpretace shluků:
(a) Prokazuje popis.
(b) Přidělení korespondence ke shlukům předvídaným z teorie.
(c) V konfirmatorním modu profily přidělují shlukům korespondenci.
(d) Při hledání korespondence nebo praktické významnosti by se měly porovnávat odvozené shluky s předem vytvořenou typologií.



6. krok: Validace a profilování shluků

Existuje subjektivní charakter hledání optimálního shlukového řešení.
Neexistuje jednoduchá metoda, která by zajišťovala validitu a praktický význam.

Validování shluků: znamená, že nalezené shlukové řešení
(a) je reprezentativní,
(b) je zobecnitelné na ostatní objekty v celém původním souboru,
(c) je stabilní i v čase.

Postup: - analyzovat oddělené výběry,
- porovnat nalezená shluková řešení a
- odhadnout shodu výsledků.

Rozdělení výběru dat na dva vzorky: každý vzorek je podroben analýze shluků odděleně a výsledky jsou porovnány:
(1) Modifikovanou formu rozdělení výběru, kdy v prvním vzorku získáme středy shluků a využijeme je k definování shluků ve druhém vzorku objektů a výsledky porovnáme,
(2) Přímá forma vzájemného porovnání (cross-validation).

Způsob vytyčení kritéria: Užijeme takové proměnné, které sice nejsou užity k vytvoření shluků, ale mění se dostatečně od shluku ke shluku.

Úloha 3. Sledování spotřeby proteinů v Evropě (Kompendium B418)

Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření.

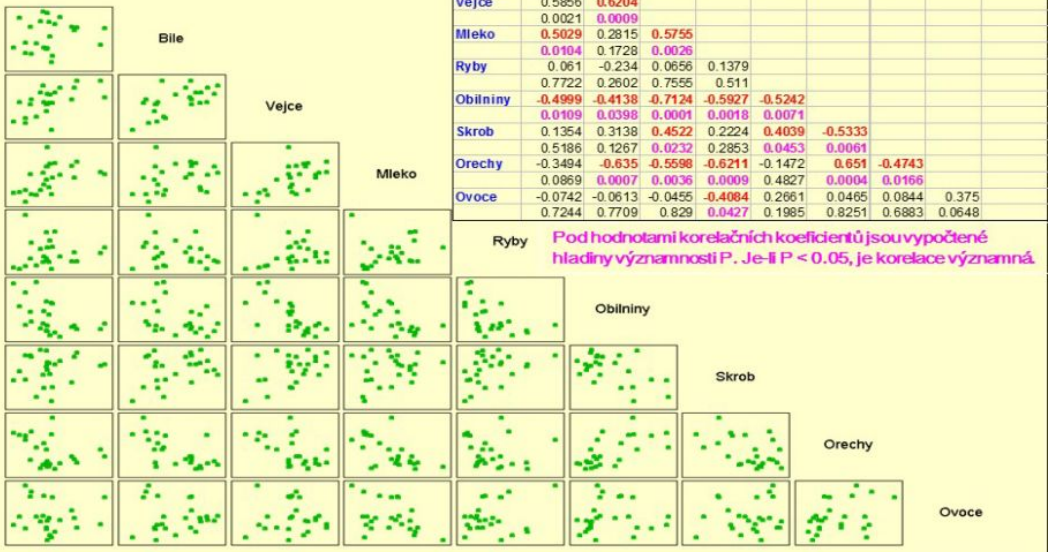
Data: *i* značí index, *Cervene* udává červené maso, **Bile** maso, **Vejce**, **Mleko**, **Ryby**, **Obilniny**, **Skrob**, **Orechy**, **Ovoce** a zelenina

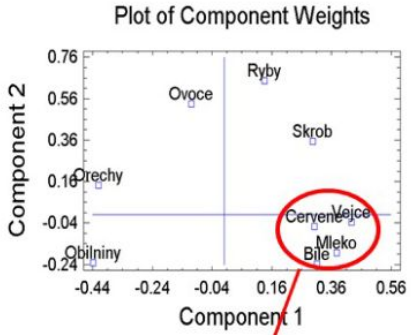
	Objekty	Proměnné								
i	Stát	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslov.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
6	Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
7	E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
8	Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
9	France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
10	Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
11	Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
12	Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
13	Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
14	Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
15	Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
16	Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
17	Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
18	Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
19	Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
20	Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
21	Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
22	UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
23	USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
24	W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

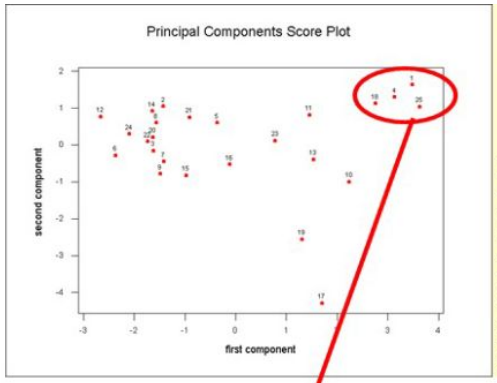
Test významnosti korelace v korelační matici

	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
Cervene									
Bile	0.153								
	0.4653								
Vejce	0.5856	0.6204							
	0.0021	0.0009							
Mleko	0.5029	0.2815	0.5755						
	0.0104	0.1728	0.0026						
Ryby	0.061	-0.234	0.0656	0.1379					
	0.7722	0.2602	0.7555	0.511					
Obilniny	-0.4999	-0.4138	-0.7124	-0.5927	-0.5242				
	0.0109	0.0398	0.0001	0.0018	0.0071				
Skrob	0.1354	0.3138	0.4522	0.2224	0.4039	-0.5333			
	0.5186	0.1267	0.0232	0.2853	0.0453	0.0061			
Orechy	-0.3494	-0.635	-0.5598	-0.6211	-0.1472	0.651	-0.4743		
	0.0869	0.0007	0.0036	0.0009	0.4827	0.0004	0.0166		
Ovoce	-0.0742	-0.0613	-0.0455	-0.4084	0.2661	0.0465	0.0844	0.375	
	0.7244	0.7709	0.829	0.0427	0.1985	0.8251	0.6883	0.0648	

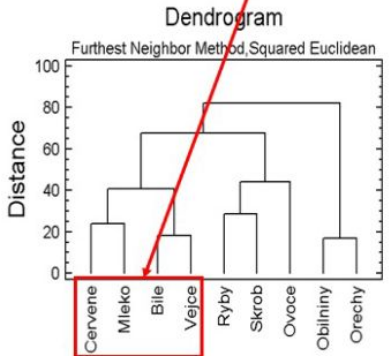
Pod hodnotami korelačních koeficientů jsou vypočtené hladiny významnosti P. Je-li $P < 0.05$, je korelace významná.

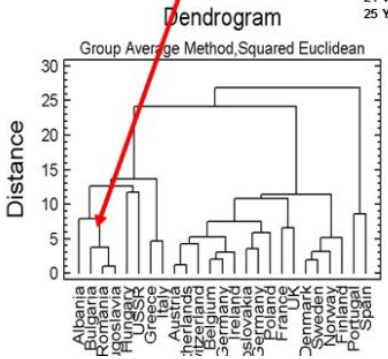






1 Albania
2 Austria
3 Belgium
4 Bulgaria
5 Czechoslovakia
6 Denmark
7 E Germany
8 Finland
9 France
10 Greece
11 Hungary
12 Ireland
13 Italy
14 Netherlands
15 Norway
16 Poland
17 Portugal
18 Romania
19 Spain
20 Sweden
21 Switzerland
22 UK
23 USSR
24 W Germany
25 Yugoslavia



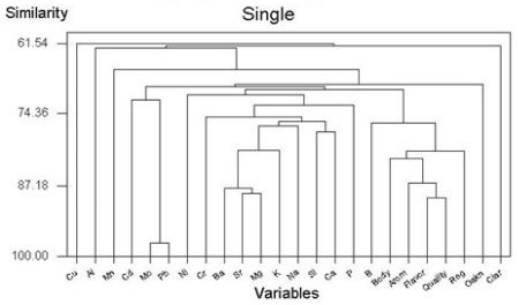


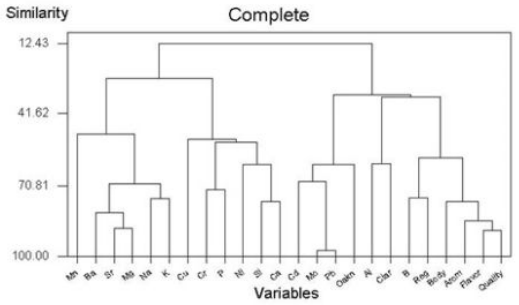
Úloha 4. Faktorová analýza při klasifikaci vzorků vín (Kompendium E408)

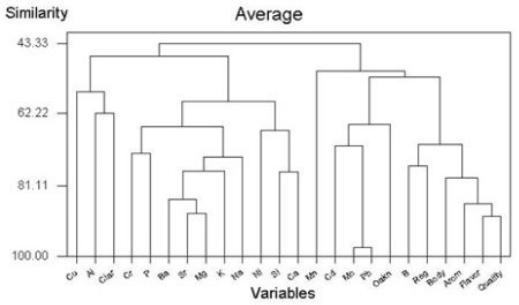
Pro 38 vzorků vín bylo nalezeno 24 analytických obsahů stopových prvků a charakteristických fyzikálně-chemických vlastností. Utvořte shluky podobných vlastností a dále shluky podobných vín.

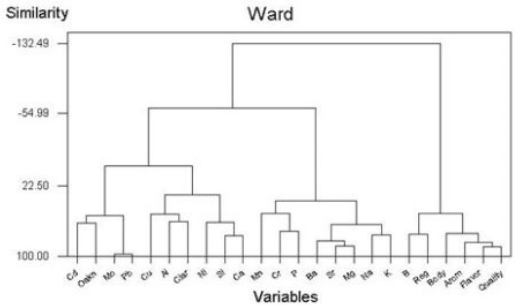
Index	Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr	Pb	B	Mg	Si	Na	Ca	P	K	Arom	Clar	Body	Flavor	Oakn	Quality	Reg
1	0.005	0.044	1.51	0.122	0.83	0.982	0.387	0.029	1.23	0.561	2.63	128	17.3	66.8	80.5	150	1130	3.3	1	2.8	3.1	4.1	9.8	1
2	0.055	0.16	1.16	0.149	0.066	1.02	0.312	0.038	0.975	0.697	6.21	193	19.7	53.3	75	118	1010	4.4	1	4.9	3.5	3.9	12.6	1
3	0.056	0.146	1.1	0.088	0.643	1.29	0.308	0.035	1.14	0.73	3.05	127	15.8	35.4	91	161	1160	3.9	1	5.3	4.8	4.7	11.9	1
4	0.063	0.191	0.959	0.38	0.133	1.05	0.165	0.036	0.927	0.796	2.57	112	13.4	27.5	93.6	120	924	3.9	1	2.6	3.1	3.6	11.1	1
5	0.011	0.363	1.38	0.16	0.051	1.32	0.38	0.059	1.13	1.73	3.07	138	16.7	76.6	84.6	164	1090	5.6	1	5.1	5.5	5.1	13.3	1
6	0.05	0.106	1.25	0.114	0.055	1.27	0.275	0.019	1.05	0.491	6.56	172	18.7	15.7	112	137	1290	4.6	1	4.7	5	4.1	12.8	1
7	0.025	0.479	1.07	0.168	0.753	0.715	0.164	0.062	0.823	2.06	4.57	179	17.8	98.5	122	184	1170	4.8	1	4.8	4.8	3.3	12.8	1
8	0.024	0.234	0.906	0.466	0.102	0.811	0.271	0.044	0.963	1.09	3.18	145	14.3	10.5	91.9	187	1020	5.3	1	4.5	4.3	5.2	12	1
9	0.009	0.058	1.84	0.042	0.17	1.8	0.225	0.022	1.13	0.048	6.13	113	13	54.4	70.2	158	1240	4.3	1	4.3	3.9	2.9	13.6	3
10	0.033	0.074	1.28	0.098	0.053	1.35	0.329	0.03	1.07	0.552	3.3	140	16.3	70.5	74.7	159	1100	4.3	1	3.9	4.7	3.9	13.9	1
11	0.039	0.071	1.19	0.043	0.163	0.971	0.105	0.028	0.491	0.31	6.56	103	9.5	45.3	67.9	133	1090	5.1	1	4.3	4.5	3.6	14.4	3
12	0.045	0.147	2.76	0.071	0.074	0.483	0.301	0.087	2.14	0.546	3.5	199	9.2	80.4	66.3	212	1470	3.3	0.5	5.4	4.3	3.6	12.3	2
13	0.06	0.116	1.15	0.055	0.18	0.912	0.166	0.041	0.578	0.518	6.43	111	11.1	59.7	83.8	139	1120	5.9	0.8	5.7	7	4.1	16.1	3
14	0.067	0.166	1.53	0.041	0.043	0.512	0.132	0.026	0.229	0.699	7.27	107	6	55.2	44.9	148	854	7.7	0.7	6.6	6.7	3.7	16.1	3
15	0.077	0.261	1.65	0.073	0.285	0.596	0.078	0.063	0.156	1.02	5.04	94.6	6.3	10.4	54.9	132	899	7.1	1	4.4	5.8	4.1	15.5	3
16	0.064	0.191	1.78	0.067	0.552	0.633	0.085	0.063	0.192	0.777	5.56	110	7	13.6	64.1	167	976	5.5	0.9	5.6	5.6	4.4	15.5	3
17	0.025	0.009	1.57	0.041	0.081	0.655	0.072	0.021	0.172	0.232	3.79	75.9	6.4	11.6	48.1	132	995	6.3	1	5.4	4.8	4.6	13.8	3
18	0.02	0.027	1.74	0.046	0.153	1.15	0.094	0.021	0.358	0.025	4.24	80.9	7.9	38.9	57.6	136	876	5	1	5.5	5.5	4.1	13.8	3
19	0.034	0.05	1.15	0.058	0.058	1.35	0.294	0.006	1.12	0.206	2.71	120	14.7	68.1	64.8	133	1050	4.6	1	4.1	4.3	3.1	11.3	1
20	0.013	0.03	2.82	0.058	0.05	0.623	0.349	0.082	2.91	0.171	3.54	208	9.3	79.2	66.4	266	1430	3.4	0.9	5	3.4	3.4	7.9	2
21	0.043	0.268	2.32	0.066	0.314	0.627	0.099	0.045	0.36	1.28	5.68	98.4	9.1	19.5	64.3	176	945	6.4	0.9	5.4	6.6	4.8	15.1	3
22	0.061	0.245	1.61	0.07	0.172	2.07	0.071	0.053	0.186	1.19	4.42	87.6	7.6	11.6	70.6	156	820	5.5	1	5.3	5.3	3.8	13.5	3
23	0.047	0.161	1.47	0.154	0.082	0.546	0.181	0.06	0.898	0.747	8.11	160	19.3	12.5	82.1	218	1220	4.7	0.7	4.1	5	3.7	10.8	2
24	0.048	0.146	1.85	0.092	0.09	0.889	0.328	0.1	1.32	0.604	6.42	134	19.3	125	83.2	173	1810	4.1	0.7	4	4.1	4	9.5	2
25	0.049	0.155	1.73	0.051	0.158	0.653	0.081	0.037	0.164	0.767	4.91	86.5	6.5	11.5	53.9	172	1020	6	1	5.4	5.7	4.7	12.7	3
26	0.042	0.126	1.7	0.112	0.21	0.508	0.299	0.054	0.995	0.686	6.94	129	43.6	45	85.9	165	1330	4.3	1	4.6	4.7	4.9	11.6	2
27	0.058	0.184	1.28	0.095	0.058	1.3	0.346	0.037	1.17	1.28	3.29	145	16.7	65.8	72.8	175	1140	3.9	1	4	5.1	5.1	11.7	1
28	0.065	0.211	1.65	0.102	0.055	0.308	0.206	0.028	0.72	1.02	6.12	99.3	27.1	20.5	95.2	194	1260	5.1	1	4.9	5	5.1	11.9	2
29	0.065	0.129	1.56	0.166	0.151	0.373	0.281	0.034	0.889	0.638	7.28	139	22.2	13.3	84.2	164	1200	3.9	1	4.4	5	4.4	10.8	2
30	0.068	0.166	3.14	0.104	0.053	0.368	0.292	0.039	1.11	0.831	4.71	125	17.6	13.9	59.5	141	1030	4.5	1	3.7	2.9	3.9	8.5	2
31	0.067	0.199	1.65	0.119	0.163	0.447	0.292	0.058	0.927	1.02	6.97	131	38.3	42.9	85.9	164	1390	5.2	1	4.3	5	6	10.7	2
32	0.084	0.266	1.28	0.087	0.071	1.14	0.158	0.049	0.794	1.3	3.77	143	19.7	39.1	128	146	1230	4.2	0.8	3.8	3	4.7	9.1	1
33	0.069	0.183	1.94	0.07	0.095	0.465	0.225	0.037	1.19	0.915	2	123	4.6	7.5	69.4	123	943	3.3	1	3.5	4.3	4.5	12.1	1
34	0.087	0.208	1.76	0.061	0.099	0.683	0.087	0.042	0.168	1.33	5.04	92.9	7	12	56.3	157	949	6.8	1	5	6	5.2	14.9	3
35	0.074	0.142	2.44	0.051	0.052	0.737	0.408	0.022	1.16	0.745	3.94	143	6.8	36.8	67.6	82	1170	5	0.8	5.7	5.5	4.8	13.5	1
36	0.084	0.171	1.85	0.088	0.038	1.21	0.263	0.072	1.35	0.899	2.38	130	6.2	101	64.4	99	1070	3.5	0.8	4.7	4.2	3.3	12.2	1
37	0.106	0.307	1.15	0.063	0.051	0.643	0.29	0.031	0.885	1.61	4.4	151	17.4	7.3	103	177	1100	4.3	0.8	5.5	3.5	5.8	10.3	1
38	0.102	0.342	4.08	0.065	0.077	0.752	0.366	0.048	1.08	1.77	3.37	145	5.3	33.1	58.3	117	1010	5.2	0.8	4.8	5.7	3.5	13.2	1

Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.

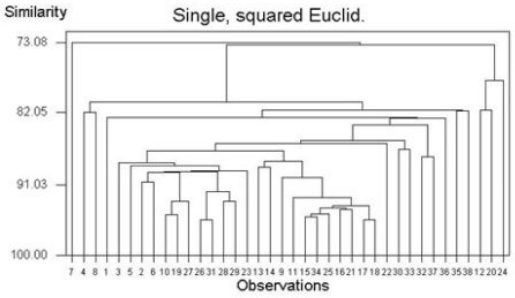


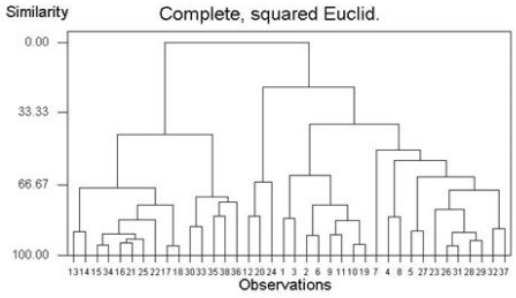


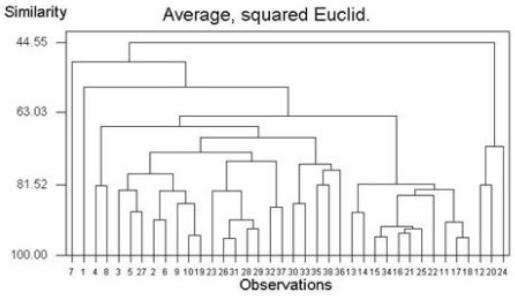


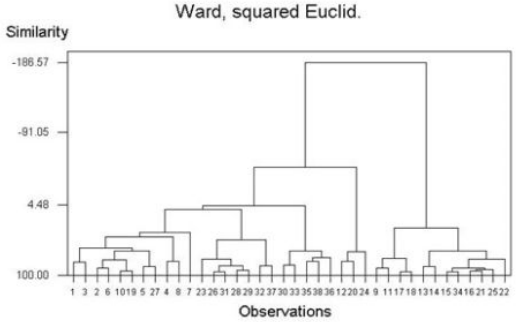


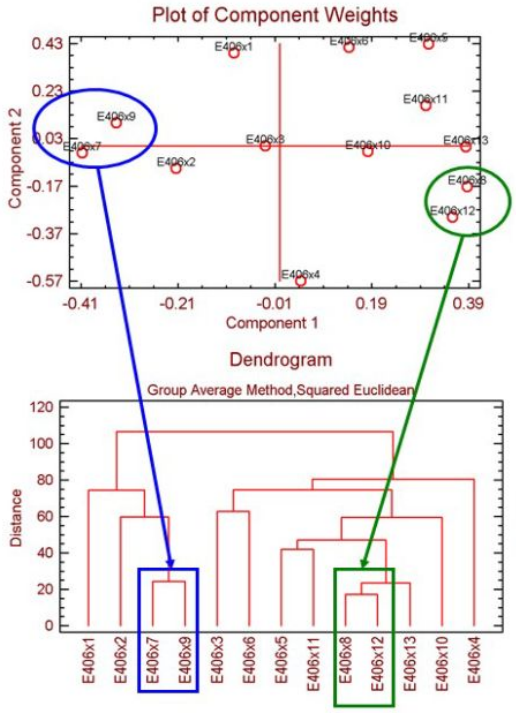
Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.

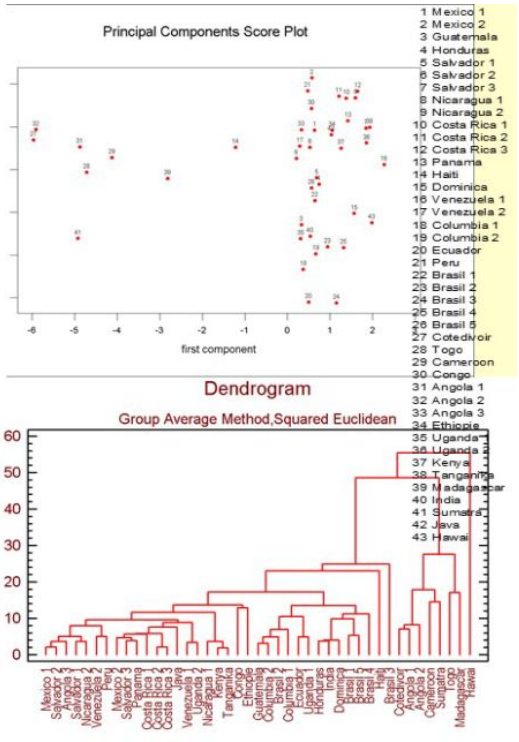












Úloha 6. Klasifikace vlastností rozličných druhů kávy (Kompendium E406)

U 43 vzorků kávy ze 30 zemí byly změřeny chemické a fyzikální vlastnosti. Nalezněte shluky podobných vlastností a shluky podobných prvků.

Data: 13 proměnných (sloupce): **i** index kávy, **j** je původ kávy, **x1** obsah vody, **x2** hmotnost zrn, **x3** extrakt, **x4** pH, **x5** volná acidita, **x6** obsah minerálů, **x7** tuky, **x8** kofein, **x9** trinonelin, **x10** kyselina chlorogeniková, **x11** kyselina neochlorogeniková, **x12** kyseliny isochlorogeniková, **x13** suma kyselin chlorogenikových.

i	ii	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1	5.4	0.4	0.8	6.6
2	Mexico 2	7.4	157.3	32.1	5.8	30.8	3.7	15	1.3	1	5.1	0.3	1	6.4
3	Guatemala	9.7	152.9	33.1	5.3	36.7	4.2	16.1	1.2	1	5.9	0.2	0.8	6.9
4	Honduras	10.4	174	31.5	5.6	34.2	3.9	15.8	1.1	0.9	5.9	0.4	0.6	6.8
5	Salvador 1	10.5	145.1	35.2	5.8	31.8	4.1	15.2	1.1	1	5.1	0.5	0.7	6.3
6	Salvador 2	10	156.4	34.5	5.8	32.6	3.9	15.4	1.2	0.8	5.3	0.4	0.7	6.4
7	Salvador 3	8.2	155.2	32.4	5.6	29.7	3.8	15.6	1.3	1.2	4.8	0.3	0.7	5.9
8	Nicaragua 1	9.2	167.8	30.6	5.9	28.9	3.8	15.1	1.3	1	5	0.3	0.7	5.9
9	Nicaragua 2	9.3	165.4	35.3	5.8	32.6	4.2	14.3	1.2	1	5.5	0.4	0.8	6.7
10	Costa Rica 1	7.1	180.3	33	5.8	29.3	4	15.1	1.3	1	5.1	0.3	0.7	6.1
11	Costa Rica 2	7.6	153.2	36	5.9	30.5	3.9	16.8	1.4	1.1	5.3	0.3	0.7	6.3
12	Costa Rica 3	7.3	159.6	35	5.8	29.9	3.7	16.5	1.2	1.2	5.5	0.3	0.7	6.5
13	Panama	9.3	161.8	32.4	5.8	31	3.7	15.5	1.3	1.2	5.6	0.3	0.6	6.6
14	Haiti	8.3	160.8	35.7	5.9	30	4.4	13	1.3	1	6.1	0.6	0.8	7.5
15	Dominica	11.6	174.8	32.5	5.4	35.2	3.7	14.5	1	1	5.7	0.3	0.5	6.5
16	Venezuela 1	9.7	169.1	34	5.8	31.6	4	15.7	1.3	1.3	5.1	0.3	0.3	6.2
17	Venezuela 2	10.6	163.7	35	5.8	35	3.8	15.8	1.2	1.1	6.1	0.3	0.9	7.3
18	Columbia 1	12	178.8	32.9	5.3	36.2	4.4	15.6	1.3	1	5.6	0.4	0.7	6.7
19	Columbia 2	10.6	169.1	33	5.3	37.5	4.4	15.1	1.2	1	6.1	0.1	0.6	6.9
20	Ecuador	11.6	148.5	34.6	5.3	39.4	4.2	14.6	1	1.1	5.7	0.5	0.4	6.6
21	Peru	10.1	153.7	34.5	6	28.4	3.7	15.9	1.3	1.1	6.1	0.4	0.8	7.3
22	Brasil 1	10.7	134.5	29.8	5.4	34.1	3.7	15.8	1.2	0.9	5.4	0.4	0.6	6.4
23	Brasil 2	9.7	160.7	33.8	5.3	37.2	4.2	15.2	1.1	0.9	5.4	0.3	0.5	6.2
24	Brasil 3	10.8	133.2	35	5.2	34.7	4.5	15.1	1.2	1.4	5	0.5	0.5	6
25	Brasil 4	11.1	131.7	29.8	5.4	33	4.1	15.8	1.1	1.2	5.1	0.5	0.5	6
26	Brasil 5	10.1	121.6	33.6	5.4	34.7	3.5	15.4	1.1	0.9	5.5	0.4	0.6	6.5
27	Cotedivoir	8	141.8	33.7	5.8	41.9	4.2	11	2	0.5	6.4	0.6	1.5	8.5
28	Togo	9	144.6	29.9	5.6	38	3.9	7.5	1.9	0.3	5.4	0.8	0.9	7.1
29	Cameroon	10.3	119.2	35.5	6.1	41.7	4.1	9.8	1.8	0.8	6	0.5	1.1	7.6
30	Congo	10	143.2	31.7	5.1	29.3	4.1	17	1.2	0.6	5.4	0.3	0.7	6.4
31	Angola 1	9.2	150.4	31.5	5.7	36.4	4.2	8.5	1.9	0.6	5.9	0.6	1.4	7.9
32	Angola 2	9.6	136.6	33.9	5.6	38.2	4	7.2	2.2	0.5	6.2	0.4	1.6	8.3
33	Angola 3	9.5	136.5	32	5.8	31.2	3.8	14.6	1.3	1	5.2	0.4	0.8	6.4
34	Ethiopie	9.3	124.2	35.6	5.8	31.8	3.8	15.7	0.9	0.9	5.5	0.2	0.8	6.5
35	Uganda 1	10.5	132.9	36.2	5.4	36.7	4	15.6	1	1	5.9	0.4	0.6	6.9
36	Uganda 2	10.7	181.2	33.1	5.8	30.7	3.9	15.8	1.3	1.1	5.3	0.3	0.6	6.2
37	Kenya	10.5	169.1	30.3	5.6	31.9	3.7	15.2	1.3	0.9	5.1	0.3	0.7	6
38	Tanganika	9.9	169.4	29	5.6	30.2	3.7	16.5	1.3	0.9	5	0.2	0.7	5.9
39	Madagascar	5	152	30.6	5.3	40.5	3.9	9.6	1.6	0.7	5.3	0.6	0.8	6.7
40	India	11.5	156.8	30.8	5.5	37.5	3.9	14.3	1.2	1	5.8	0.4	0.4	6.6
41	Sumatra	8.4	110.8	31.6	5.7	43.4	4.5	10.1	1.7	0.8	6.3	0.7	0.9	7.9
42	Java	8.6	163.1	34.5	5.5	33.3	4	16	1.2	1.1	5.1	0.3	0.8	6.3
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.8	0.3	6.8

Postup analýzy vícerozměrných dat

1. Standardizace: analýze vždy předchází standardizace čili škálování proměnných.

2. Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:
Vyčíslení výběrové střední hodnoty každé proměnné.

Odhad kovarianční matice ***S*** a její normované podoby - korelační matice ***R***.

Odhadu vícerozměrné šikmosti a vícerozměrné špičatosti.

Matice ***R*** obsahuje Pearsonovy párové korelační koeficienty, které se diskutují.

3. Exploratorní analýza dat EDA:

(a) Hledání podobnosti objektů vizuálními rozptylovými diagramy typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy),

(b) Nalezení vybočujících objektů nebo vybočujících proměnných, mnohdy nevhodných k analýze,

(c) Testy předpokladů lineárních vazeb,

(d) Testy předpokladů o datech (normalitu, nekorelovanost, homogenitu).
Ověřování normality založené na vícerozměrné šikmosti a vícerozměrné špičatosti.

4. Určení vhodného počtu latentních proměnných:

a) Matice ***S*** nebo ***R*** se rozloží na vlastní čísla a vlastní vektory.

b) Indexový graf úpatí vlastních čísel (Scree plot): určí vhodný počet latentních proměnných, které ještě dostatečně popisují proměnlivost v datech.

c) Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. Jinak jde o hlavní komponenty.

5. Určení struktury v proměnných (PCA a FA):

a) Graf komponentních vah (Plot of components weights, loadings): hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu

b) Rozptylový diagram komponentního skóre (Scatterplot): hledání struktury v objektech a třídění objektů do shluků.

c) Dvojný graf (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

6. Určení struktury a vzájemných vazeb v objektech:

a) Klasifikační postupy zařadí analyzovaný objekt do jednoho již existujícího a předem zadaného shluku.

b) Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu.

c) Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístěni typičtí reprezentanti.

7. Vysvětlení souladu nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA (či FA) grafech:

a) Vyšetřit a vysvětlit nalezenou strukturu a vazby jednotlivých proměnných nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků.

b) Vysvětlit strukturu a vazby klasifikovaných objektů nalezenou v PCA a v dendrogramu podobnosti objektů.

Řešené úlohy

PŘÍKLAD 9.4 *Vytvoření dendrogramu neuroleptik*

Neuroleptika redukují nežádoucí účinky přebytečného dopaminu a liší se ve svých účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění atd. Cílem je provést klasifikaci neuroleptik do shluků podobných účinků.

○ ***Data*: Data *Neuroleptika*** (převrácená hodnota mediánové účinné dávky *1/ED50* [kg/mg]):
Lek název neuroleptika,
Nervoz potlačení nervozity,
Stereo potlačení stereotypního chování,
Tres potlačení záchvatu a třesu a
Usmr dávka smrtícího účinku.

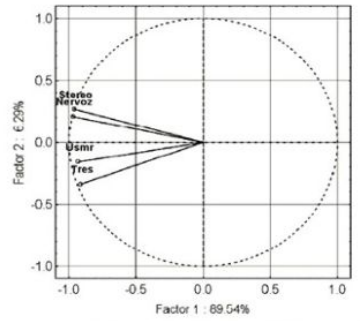
Lek	*Nervoz*	*Stereo*	*Tres*	*Usmr*
1 Chlorphromazine	3.846	3.333	1.111	1.923
2 Promazine	0.323	0.213	0.108	1.429
3 Trifluperazine	27.027	17.857	0.562	0.14
4 Fluphenazine	17.857	15.385	1.695	1.075
5 Perphenazine	27.027	27.027	1.961	2.083
6 Thioridazine	0.244	0.185	0.093	1.333
7 Pifluthixol	142.857	142.857	20.408	163.934
8 Thiothixene	4.348	4.348	0.047	0.345
9 Chorprothixene	5.882	2.941	4.545	4.167
10 Spiperone	62.5	47.619	11.765	0.847
11 Haloperidol	52.632	62.5	1.282	0.568
12 Azaperone	2.941	1.282	2.222	3.03
13 Pipamperone	0.327	0.187	1.724	0.397
14 Pimozide	20.408	20.408	0.107	0.025
15 Metitepine	15.385	10.204	10.204	27.027
16 Clozapine	0.161	0.093	0.327	0.323
17 Perlapine	0.323	0.323	0.37	0.067
18 Sulpiride	0.047	0.047	0.003	0.001
19 Butaclamol	10.204	9.091	1.471	0.025
20 Molindone	7.692	7.692	0.14	38138

○ ***Řešení:*** Po vyhledání optimální tvorby dendrogramu sestrojíme dendrogram podobnosti znaků a dendrogram podobnosti objektů.

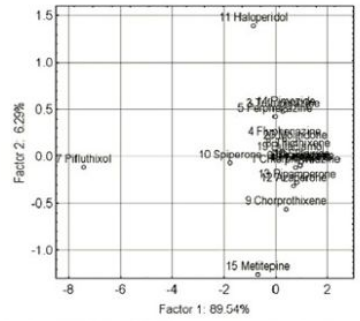
Nejvyšší hodnota kofenetického korelačního koeficientu ***CC*** a nejnižší hodnota obou kritérií delta, ***Delta(0.5)*** a ***Delta(1.0)***, vybrala **metodu skupinového průměru** (software NCSS2004).

1. Nejbližšího souseda, *Kofenetická korelace CC:* 0.988598, *Delta(0.5):* 0.474238, *Delta(1.0):* 0.391993.
2. Nejvzdálenějšího souseda: *Kofenetická korelace CC:* 0.982795, *Delta(0.5):* 0.178589, *Delta(1.0):* 0.183477;
3. Párový průměr, *Kofenetická korelace CC:* 0.988876, *Delta(0.5):* 0.177810, *Delta(1.0):* 0.188781;
4. **Skupinový průměr**, *Kofenetická korelace CC:* 0.987356, *Delta(0.5):* 0.137455, *Delta(1.0):* 0.125290;
5. Těžiště, *Kofenetická korelace CC:* 0.984750, *Delta(0.5):* 0.175238, *Delta(1.0):* 0.166599;
6. Median, *Kofenetická korelace CC:* 0.984215, *Delta(0.5):* 0.452308, *Delta(1.0):* 0.428346;
7. Wardova metoda, *Kofenetická korelace CC:* 0.979285, *Delta(0.5):* 0.549394, *Delta(1.0):* 0.492716.

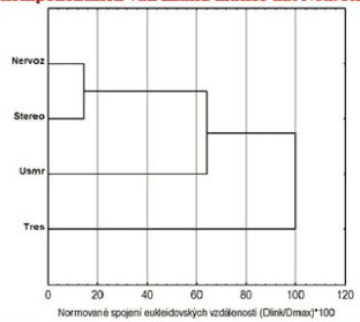
Metoda skupinového průměru v dendrogramu podobnosti objektů:
první shluk obsahuje 12 objektů 1, 8, 12, 9, 2, 6, 16, 17, 18, 13, 19, 20,
druhý shluk 5 objektů 3, 4, 14, 5, 15,
třetí shluk 2 objekty 10 a 11,
čtvrtý shluk obsahuje jeden objekt, a to 7.



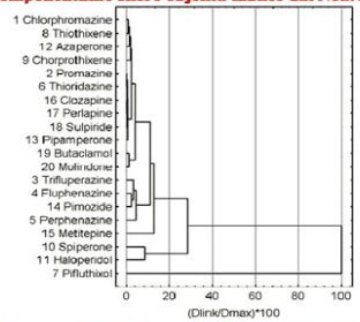
Graf komponentních vah znaků matice dat *Neuroleptika*.



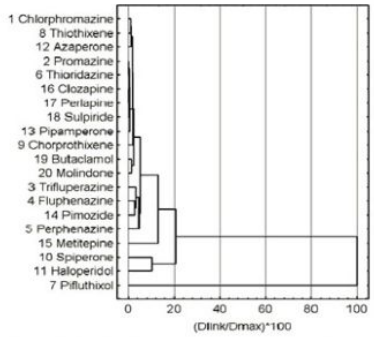
Graf komponentního skóre objektů matice dat *Neuroleptika*



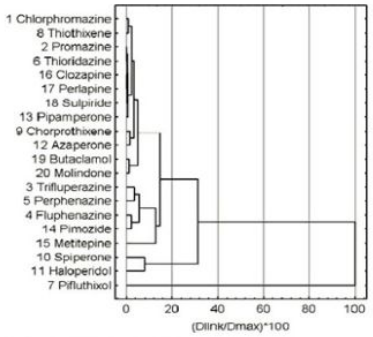
Dendrogram znaků metodou skupinového průměru



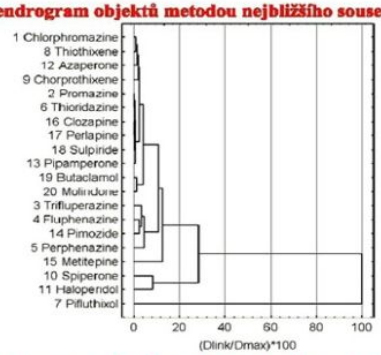
Dendrogram objektů metodou skupinového průměru



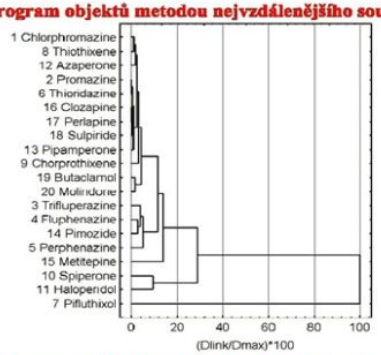
Dendrogram objektů metodou nejbližšího souseda



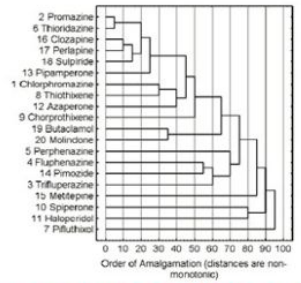
Dendrogram objektů metodou nejvzdálenějšího souseda



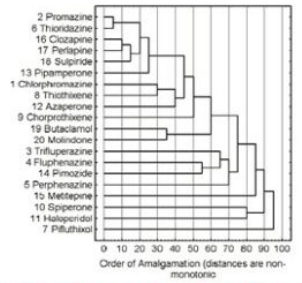
Dendrogram objektů metodou párového průměru



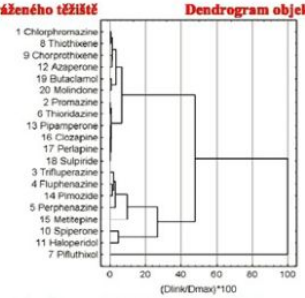
Dendrogram objektů metodou skupinového průměru



Dendrogram objektů metodou neváženého těžiště

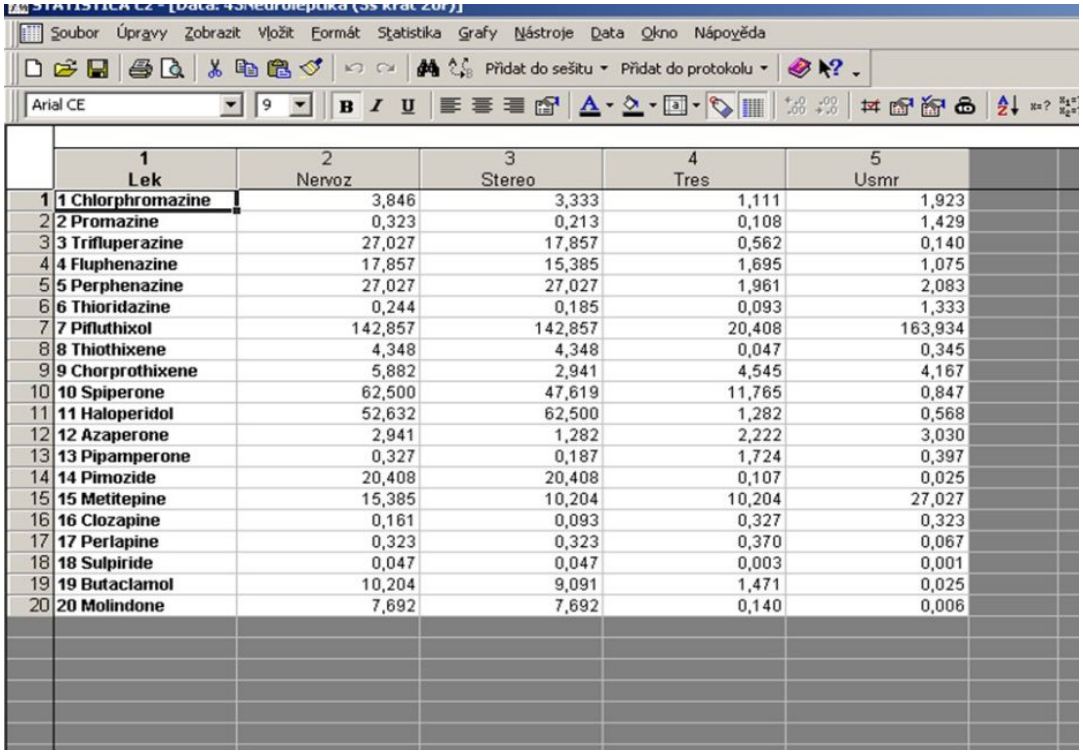


Dendrogram objektů metodou váženého těžiště (mediánu).

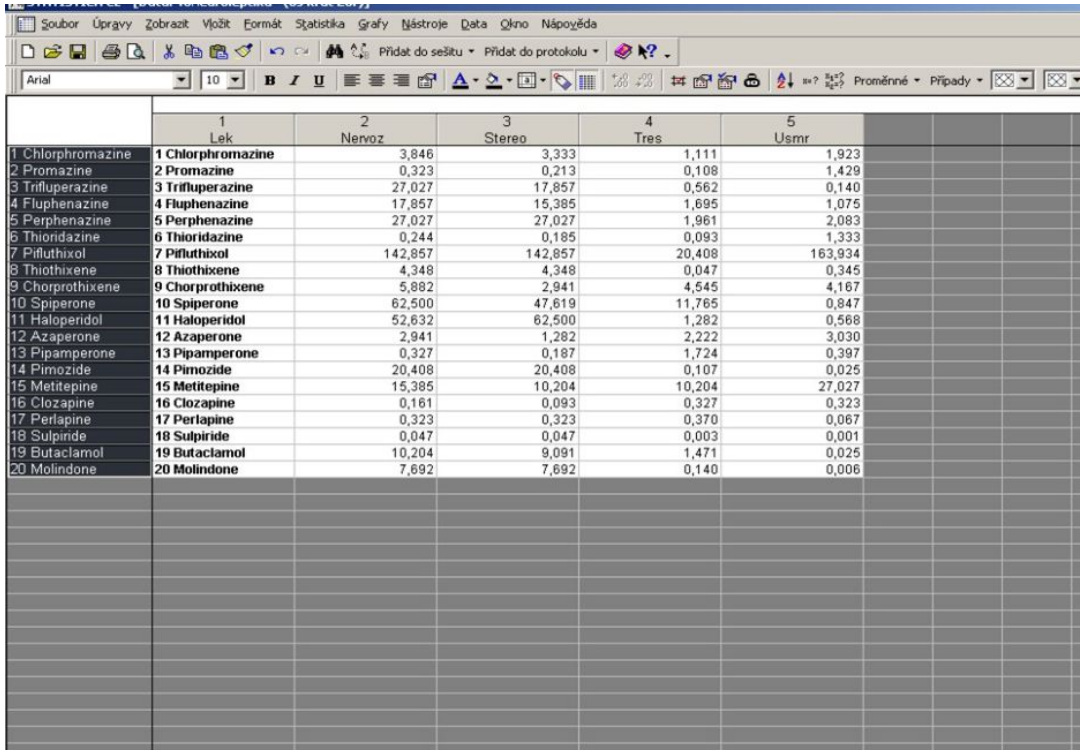


Dendrogram objektů metodou Wardovou

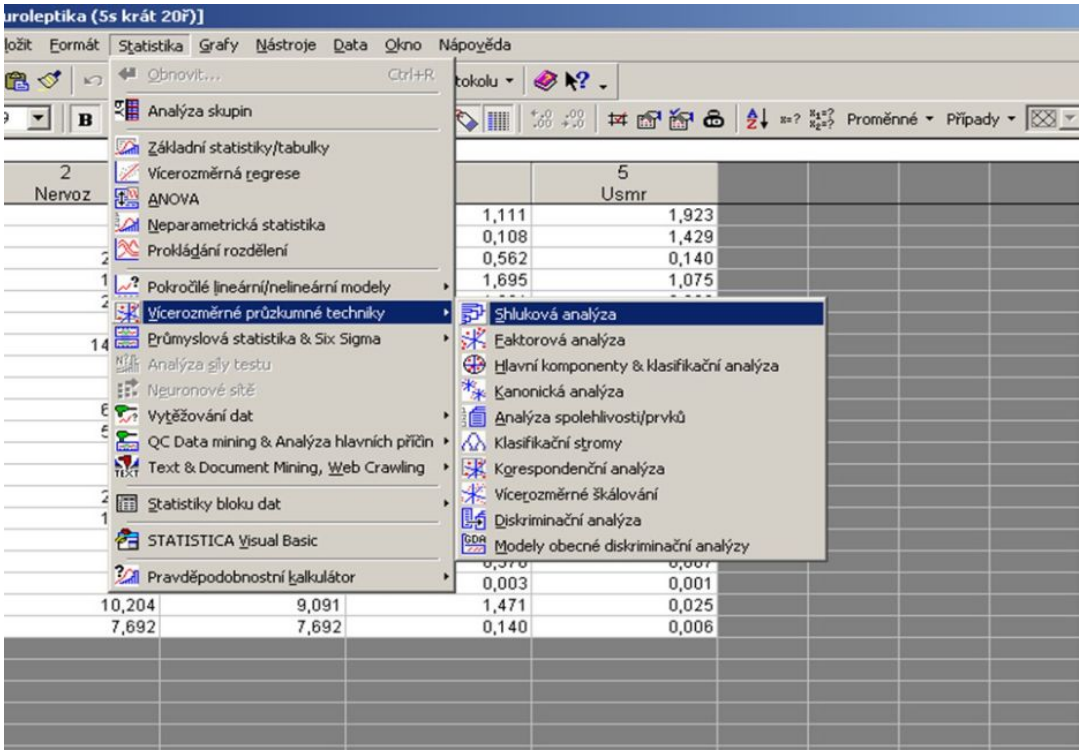
Závěr: Nejvhodnější tvorba dendrogramu je metodami párového průměru a skupinového průměru.

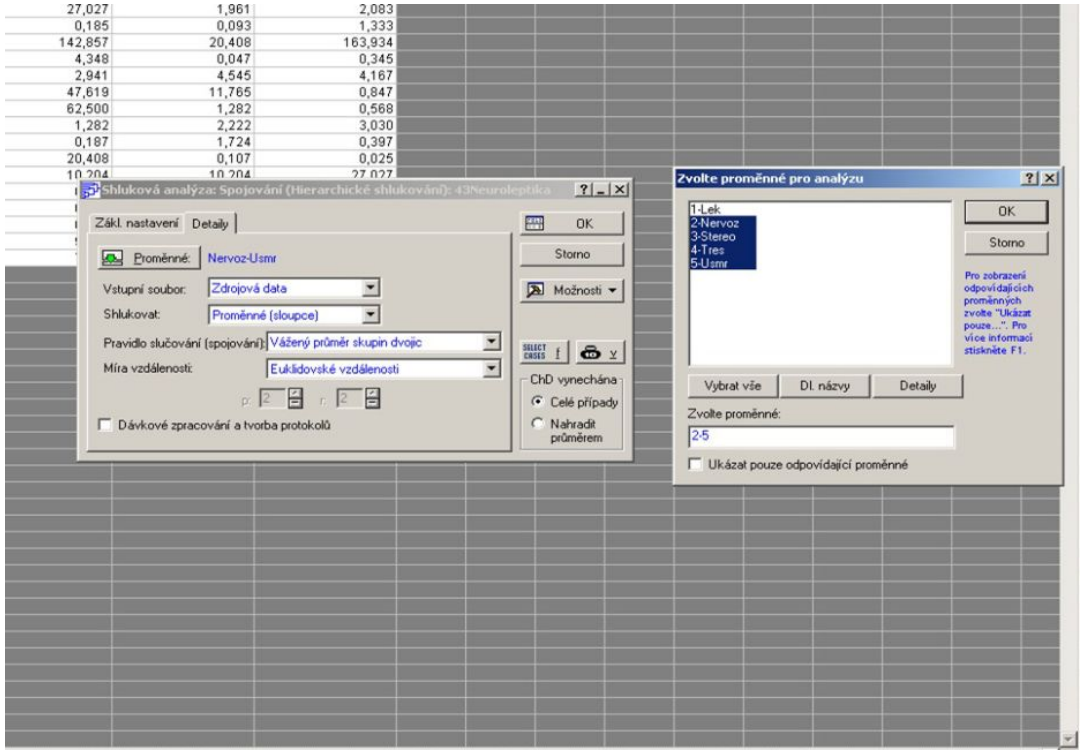


	1 Lek	2 Nervoz	3 Stereo	4 Tres	5 Usmr
1	1 Chlorphromazine	3,846	3,333	1,111	1,923
2	2 Promazine	0,323	0,213	0,108	1,429
3	3 Trifluperazine	27,027	17,857	0,562	0,140
4	4 Fluphenazine	17,857	15,385	1,695	1,075
5	5 Perphenazine	27,027	27,027	1,961	2,083
6	6 Thioridazine	0,244	0,185	0,093	1,333
7	7 Pifluthixol	142,857	142,857	20,408	163,934
8	8 Thiothixene	4,348	4,348	0,047	0,345
9	9 Chorprothixene	5,882	2,941	4,545	4,167
10	10 Spiperone	62,500	47,619	11,765	0,847
11	11 Haloperidol	52,632	62,500	1,282	0,568
12	12 Azaperone	2,941	1,282	2,222	3,030
13	13 Pipamperone	0,327	0,187	1,724	0,397
14	14 Pimozide	20,408	20,408	0,107	0,025
15	15 Metitepine	15,385	10,204	10,204	27,027
16	16 Clozapine	0,161	0,093	0,327	0,323
17	17 Perlapine	0,323	0,323	0,370	0,067
18	18 Sulpiride	0,047	0,047	0,003	0,001
19	19 Butaclamol	10,204	9,091	1,471	0,025
20	20 Molindone	7,692	7,692	0,140	0,006



	1 Lek	2 Nervoz	3 Stereo	4 Tres	5 Usmr
1 Chlorphromazine	1 Chlorphromazine	3,846	3,333	1,111	1,923
2 Promazine	2 Promazine	0,323	0,213	0,108	1,429
3 Trifluperazine	3 Trifluperazine	27,027	17,857	0,562	0,140
4 Fluphenazine	4 Fluphenazine	17,857	15,385	1,695	1,075
5 Perphenazine	5 Perphenazine	27,027	27,027	1,961	2,083
6 Thioridazine	6 Thioridazine	0,244	0,185	0,093	1,333
7 Pifluthixol	7 Pifluthixol	142,857	142,857	20,408	163,934
8 Thiothixene	8 Thiothixene	4,348	4,348	0,047	0,345
9 Chorprothixene	9 Chorprothixene	5,882	2,941	4,545	4,167
10 Spiperone	10 Spiperone	62,500	47,619	11,765	0,847
11 Haloperidol	11 Haloperidol	52,632	62,500	1,282	0,568
12 Azaperone	12 Azaperone	2,941	1,282	2,222	3,030
13 Pipamperone	13 Pipamperone	0,327	0,187	1,724	0,397
14 Pimozide	14 Pimozide	20,408	20,408	0,107	0,025
15 Metitepine	15 Metitepine	15,385	10,204	10,204	27,027
16 Clozapine	16 Clozapine	0,161	0,093	0,327	0,323
17 Perlapine	17 Perlapine	0,323	0,323	0,370	0,067
18 Sulpiride	18 Sulpiride	0,047	0,047	0,003	0,001
19 Butaclamol	19 Butaclamol	10,204	9,091	1,471	0,025
20 Molindone	20 Molindone	7,692	7,692	0,140	0,006





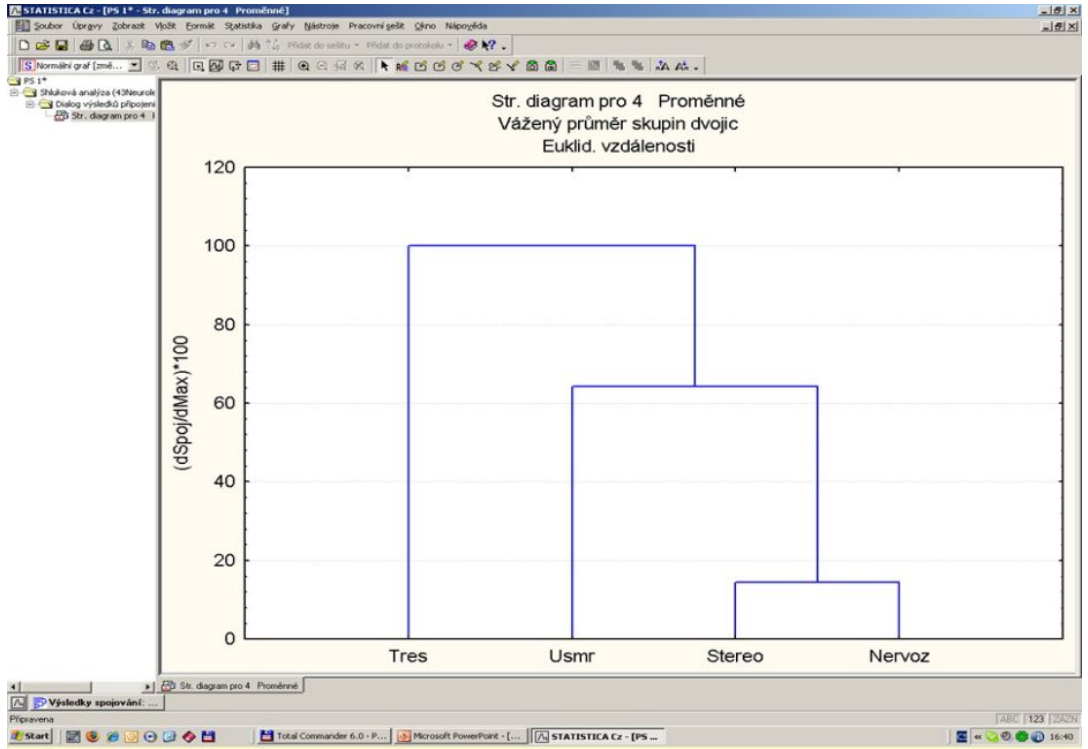
,857	0,562	0,140
,385	1,695	1,075
,027	1,961	2,083
),185	0,093	1,333
,857	20,408	163,934
,348	0,047	0,345
,941	4,646	4,167
,619		
,500		
,282		
),187		
),408		
),204		
),093		
),323		
),047		
),091		
,692		

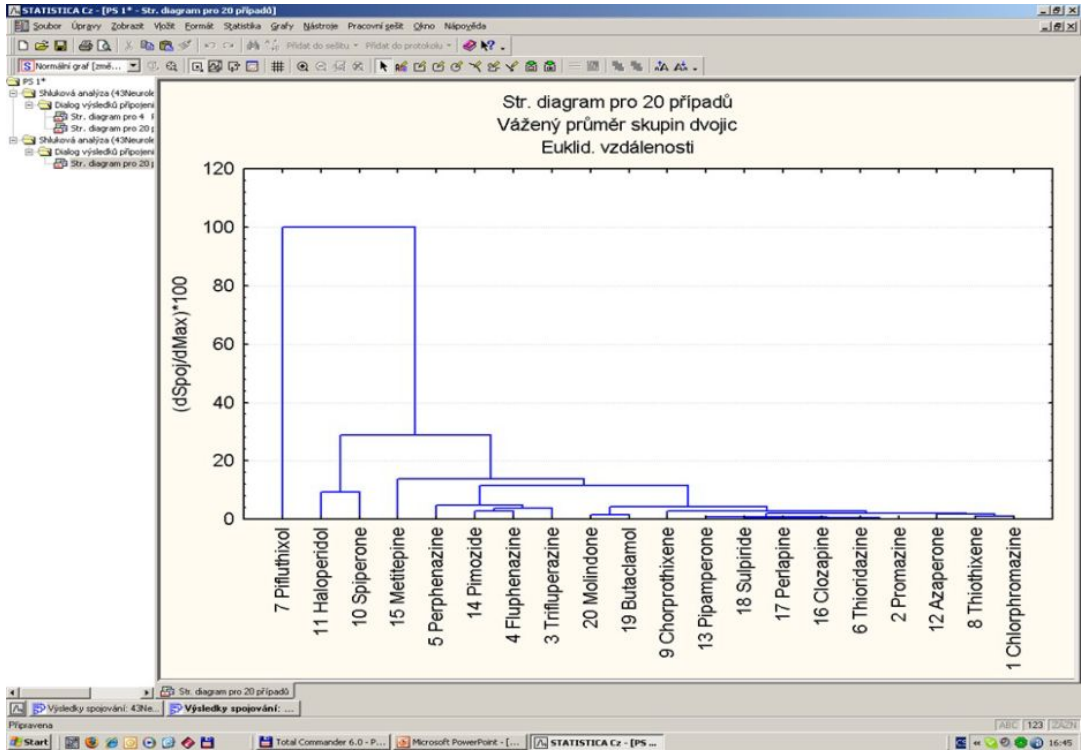
Výsledky spojování: 43Neuroleptika

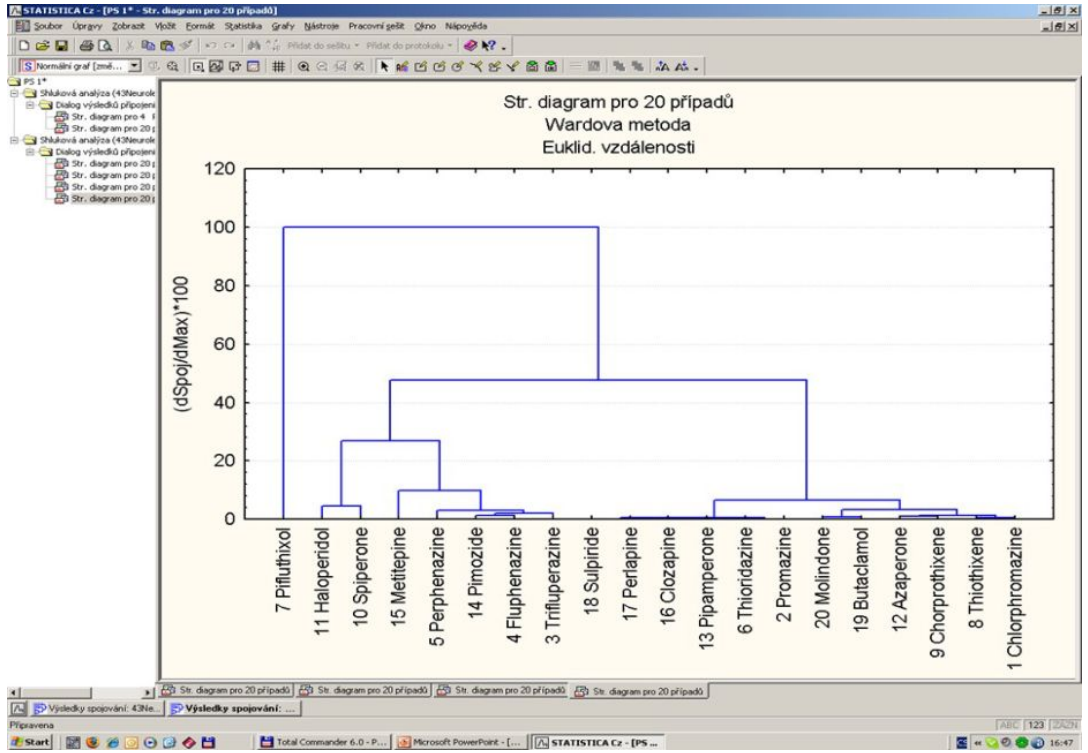
Počet proměnných: 4
Počet případů: 20
Spojování proměnných
Chybějící data odstr. případově
Pravidlo slučování (spoj.): Vážený průměr skupin dvojic
Metrika vzdálenosti: Euklid. vzdálenosti (nestandardiz.)

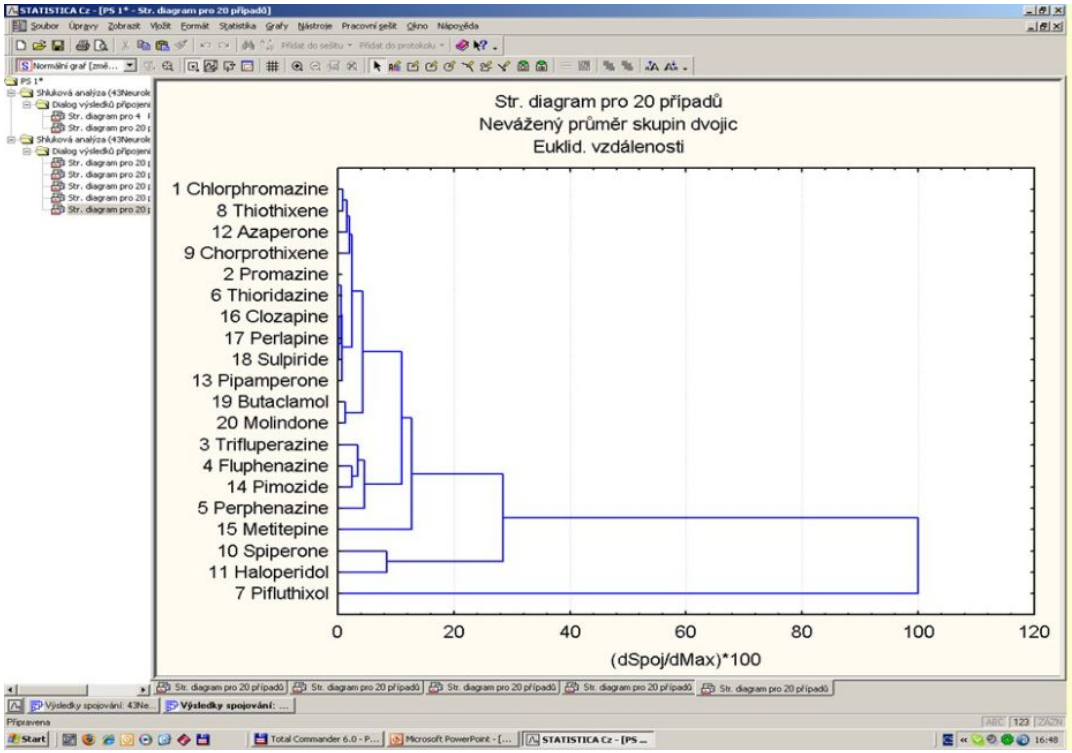
Zákl. výsledky | Detaily

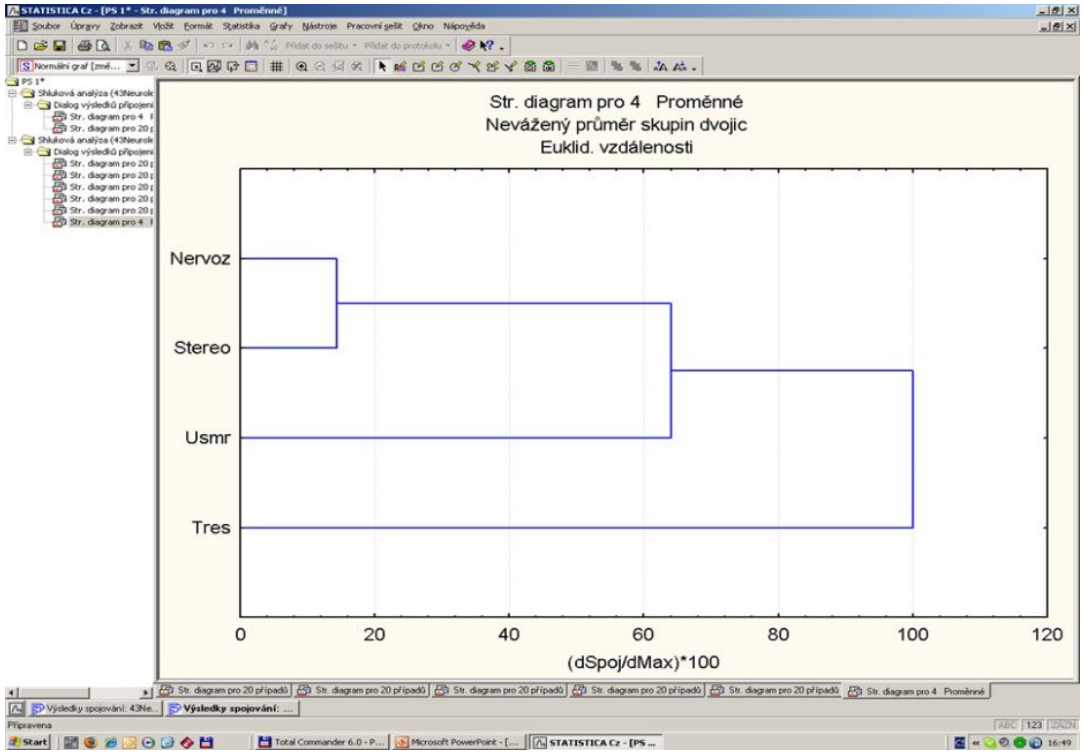
Horizontální graf hierarch. stromu
Vertikální "třásňový" graf
Pravoúhlé větve
Standardizovat měřítko stromu (*100)
Rozvrh shlukování
Graf rozvrhu shlukování
Matice vzdáleností
Popisné statistiky
Matice
Souhrn
Storno
Možnosti

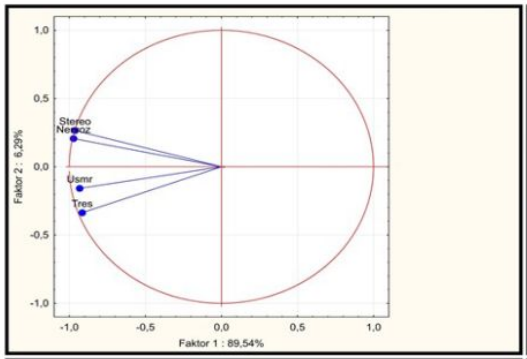


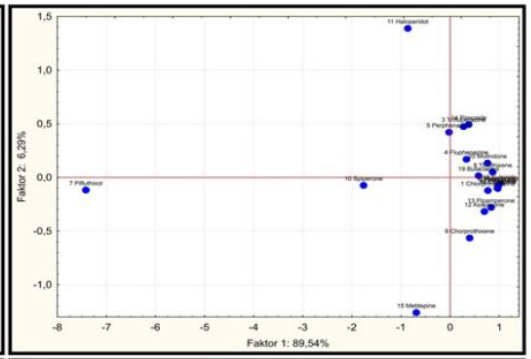


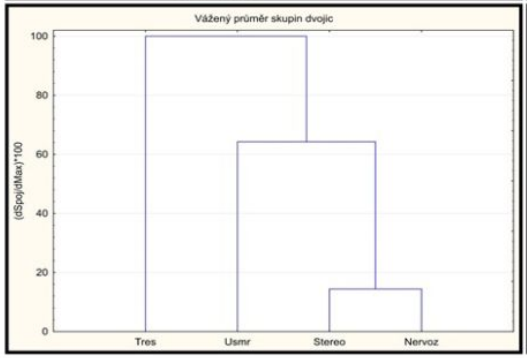


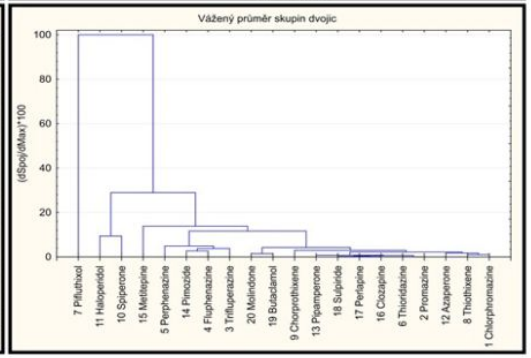












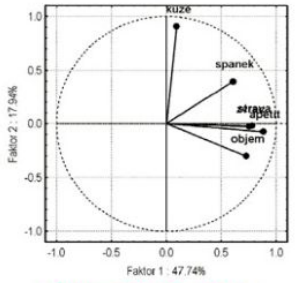
PŘÍKLAD 9.11 *Výstavba shluků u radioterapeutického léčení vybraných pacientů*

U 98 pacientů byl sledováno radioterapeutické léčení. Do kolika shluků se roztřídí 98 pacientů?

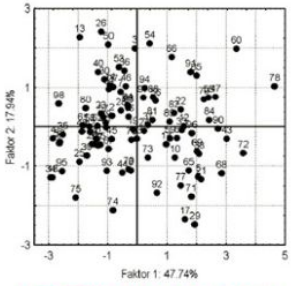
○ ***Data*:** Data *Radioterapie* obsahuje 98 pacientů 6 sledovaných znaků:
Pacient je index pacienta,
Zvrac počet symptomů jako je pálení žáhy, zvracení atd.,
Objem značí objem provedených činností ve stupnici 1 až 5,
Spanek značí objem spánku ve stupnici 1 až 5,
Strava značí množství zkonzumované stravy,
Apetit značí apetit ve stupnici 1 až 5,
Kuze značí podrážděnost kůže ve stupnici 0 až 3.

Pacient	*Zvrac*	*Objem*	*Spanek*	*Strava*	*Apetit*	*Kuze*
1	0.889	1.389	1.555	2.222	1.945	1
...	...	...	...	...		...
98	0.889	1	1	2	1	2

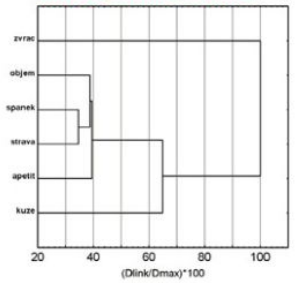
○ ***Řešení:*** Graf komponentních vah znaků ukazuje silnou korelaci znaků ***Objem, Apetit, Strava*** a ***Zvrac***, protože tyto čtyři znaky jsou v grafu představeny téměř totožnými průvodiči.



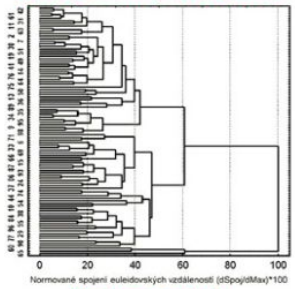
Graf komponentních vah znaků.



Graf komponentního skóre 98 pacientů

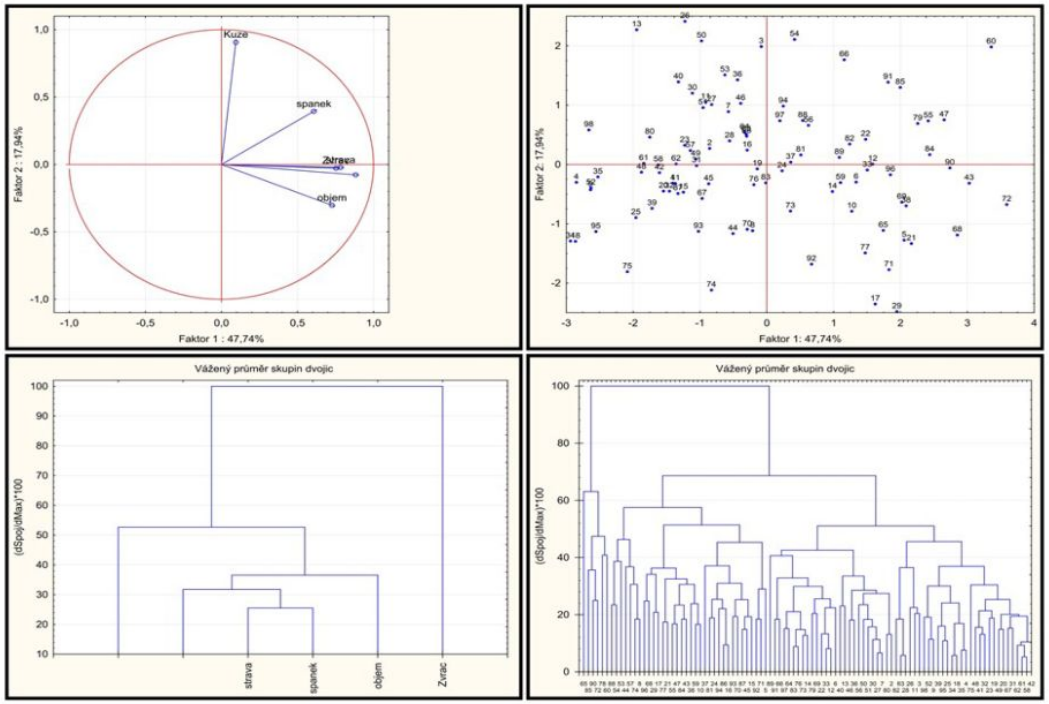


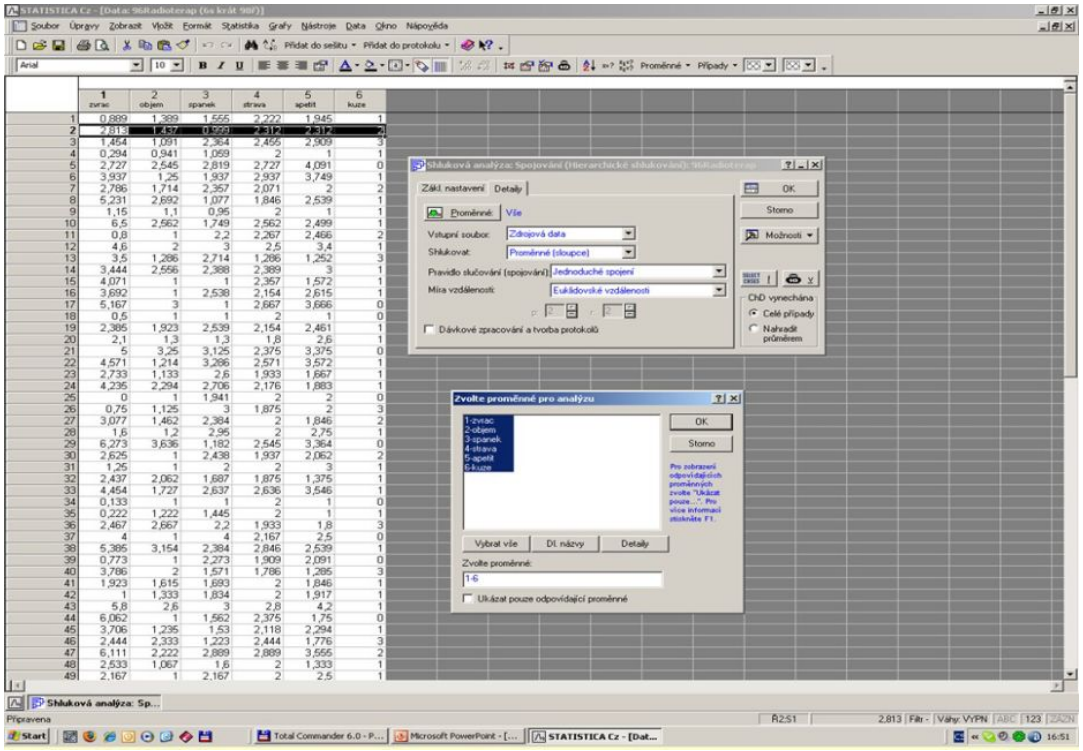
Dendrogram znaků

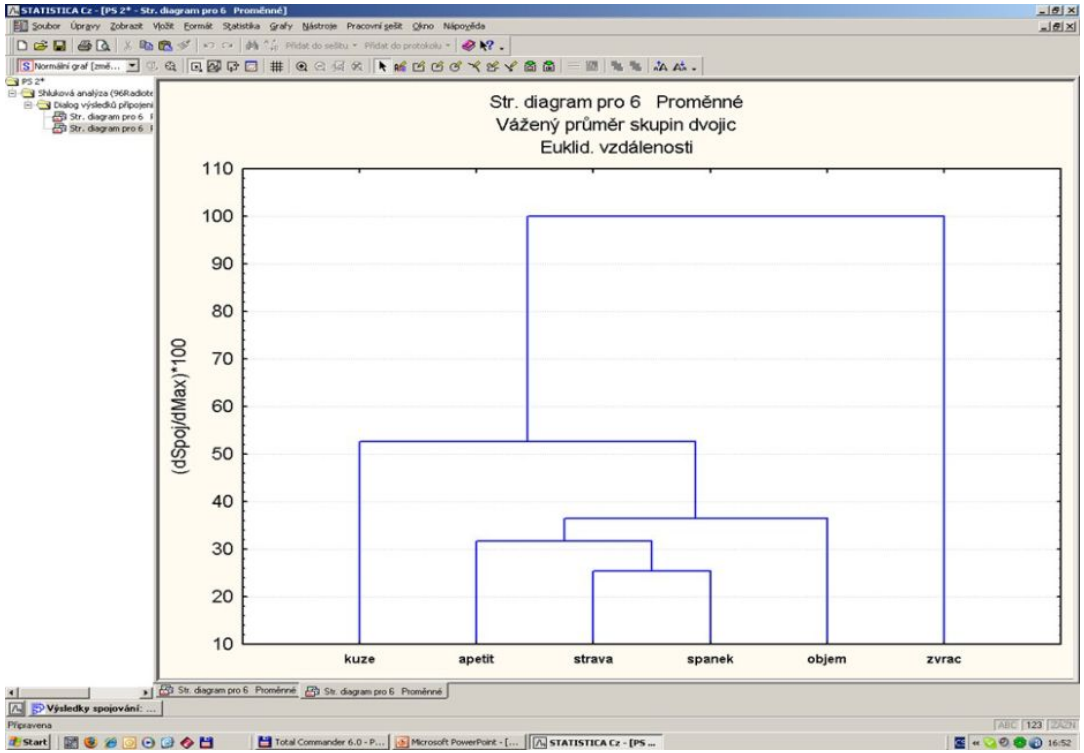


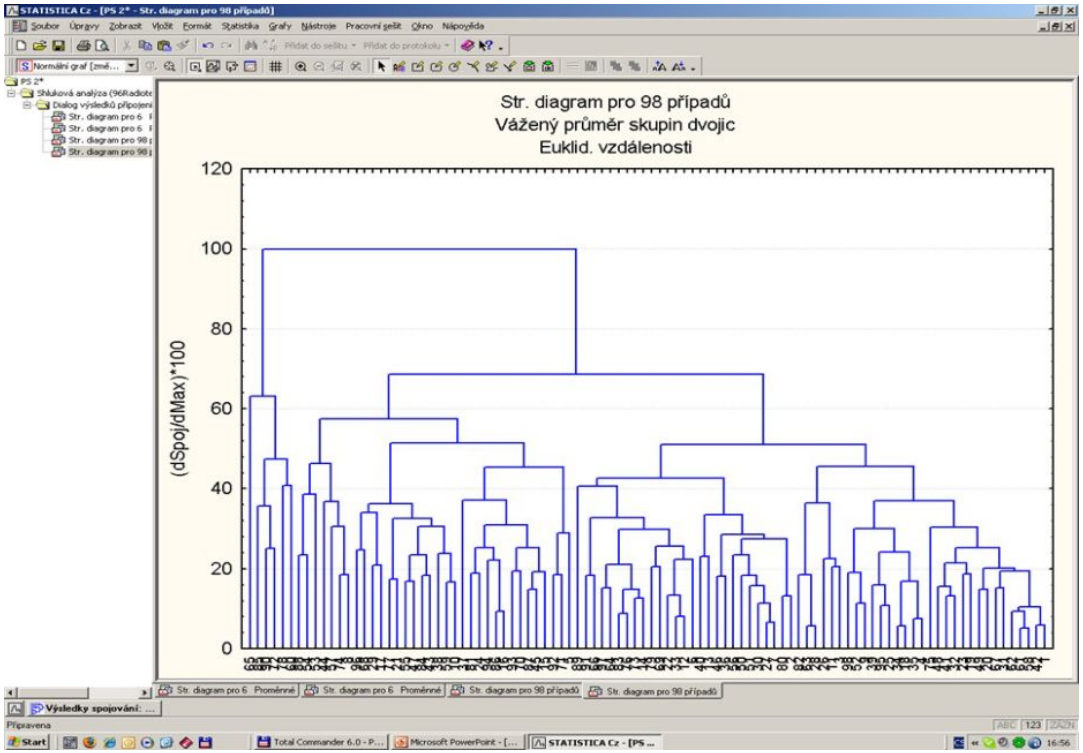
Dendrogram pacientů

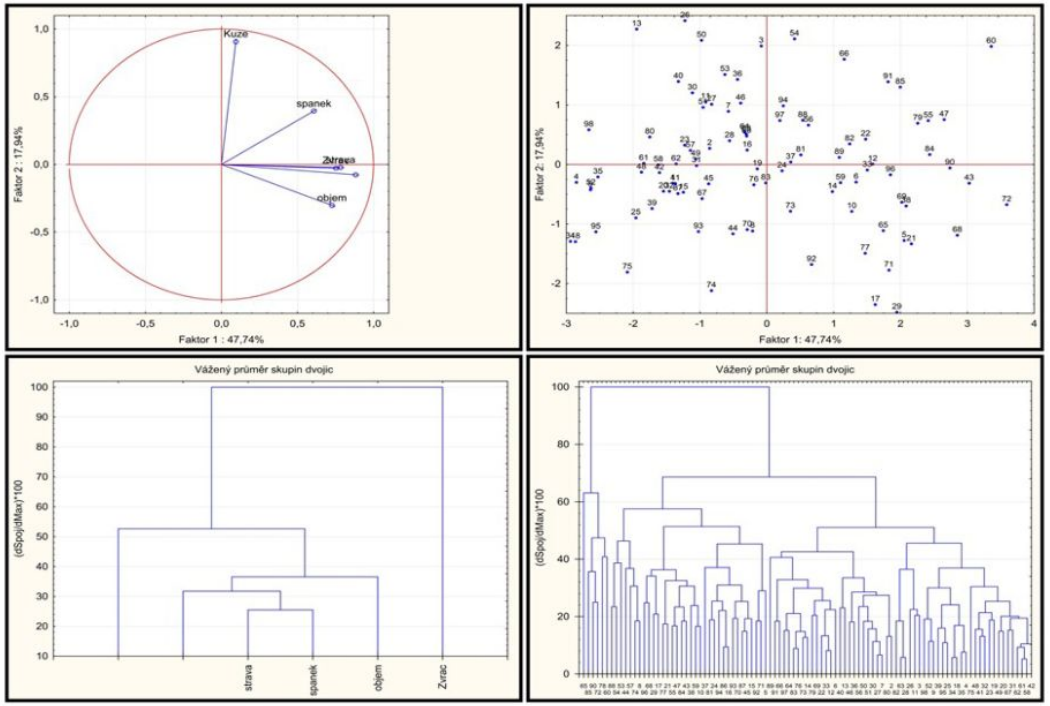
Závěr: Dendrogram objektů klasifikuje 98 pacientů do několika shluků a 3 pacienti jsou odlišní.











PŘÍKLAD 9.12 *Dendrogram úbytku kostní hmoty starších žen po cvičeních a po dietách*

Zkoumáno, zda cvičení nebo doplňky vhodné diety zpomalí úbytek kostní hmoty u žen. Obsah minerálů v kostech byl měřen absorpční fotometrií ve třech kostech na dominantní a ve třech na vedlejší straně. Při klasifikaci je třeba sestrojit dendrogram blízkých znaků a dendrogram vzniklých shluků pacientů.

○ ***Data:*** Data *Kost* obsahuje 25 pacientů obsah minerálů v 6 vyšetřovaných znacích:
Pacient je index pacienta,
Domin značí poloměr u dominantní kosti,
Vedlej značí poloměr u vedlejší kosti,
Dopaze značí dominantní část kosti pažní,
Vepaze značí vedlejší část kosti pažní,
Doloket značí dominantní část kosti loketní a
Veloket značí vedlejší část kosti loketní.

Pacient	*Domin*	*Vedlej*	*Dopaze*	*Vepaze*	*Doloket*	*Veloket*
1	1.103	1.052	2.139	2.238	0.873	0.872
...	...	...	...	...		...
25	0.915	0.936	1.971	1.869	0.869	0.868

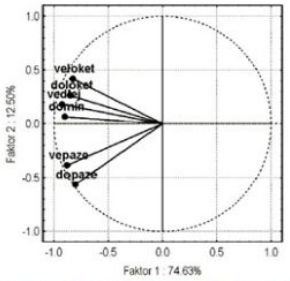
○ ***Řešení:*** **Graf komponentních vah znaků** ukazuje silnou korelaci a podobnost dvojic znaků *Domin-Vedlej*, dále *Doloket-Veloket* a konečně také *Dopaze-Vepaze.*

Dvě dvojice *Domin-Vedlej* a *Doloket-Veloket* spolu rovněž korelují a dle polohy v grafu jsou si také podobné.

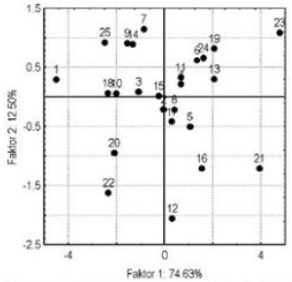
Dendrogram znaků ukazuje ve shodě s předešlým grafem na vznik dvou blízkých shluků, první obsahuje znaky *Domin* a *Vedlej* a druhý shluk obsahuje *Doloket* a *Veloket*, který je méně podobný třetímu shluku, který obsahuje dvojici *Dopaze* a *Vepaze.*

Umístění pacientů na grafu komponentního skóre objektů je vcelku ve shodě s dendrogramem pacientů.

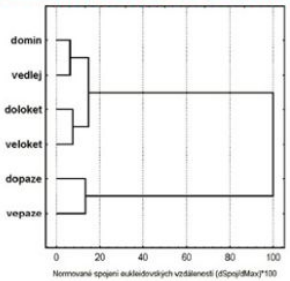
Lze indikovat tři shluky:
První obsahuje objekty 1, 20, 22, 10, 18, 25 a 12.
Druhý velký shluk obsahuje 2, 5, 8, 16, 17, 4, 11, 3, 9, 14, 7 a 15.
Třetí shluk obsahuje objekty 6, 13, 24, 19, 21.
Objekt 23 je odlehlý, nepodobný všem ostatním.



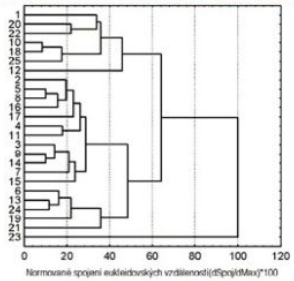
Graf komponentních vah znaků matice dat *Kost*, (STATISTICA).



Graf komponentního skóre 25 pacientů matice dat *Kost*

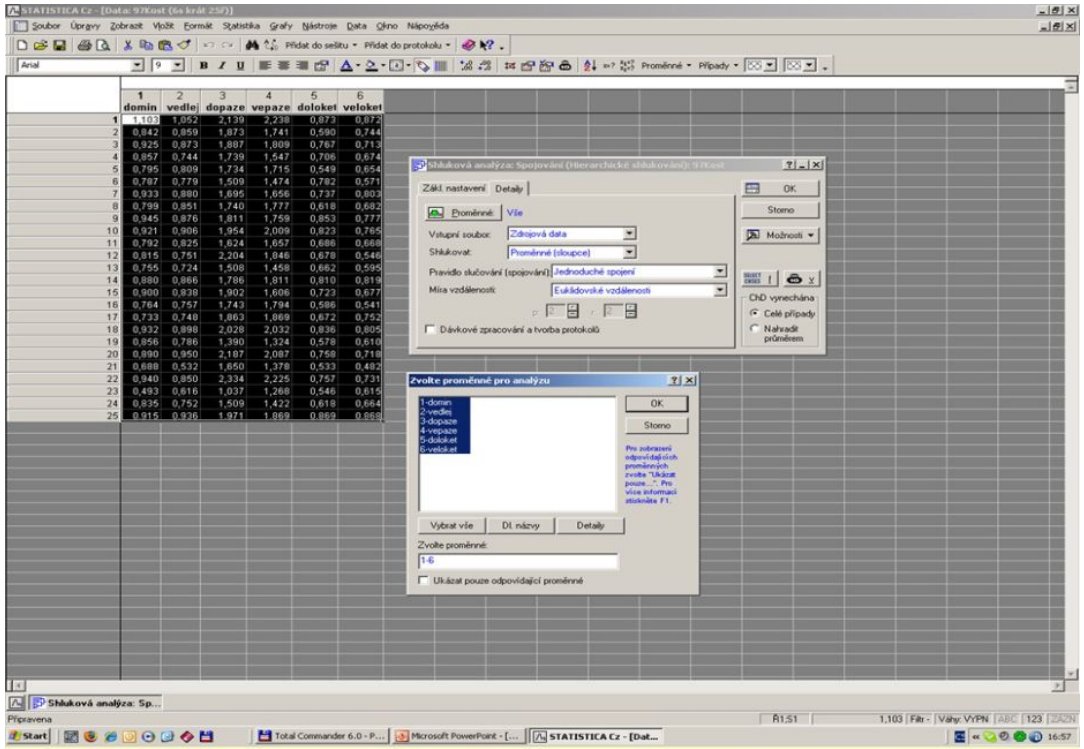


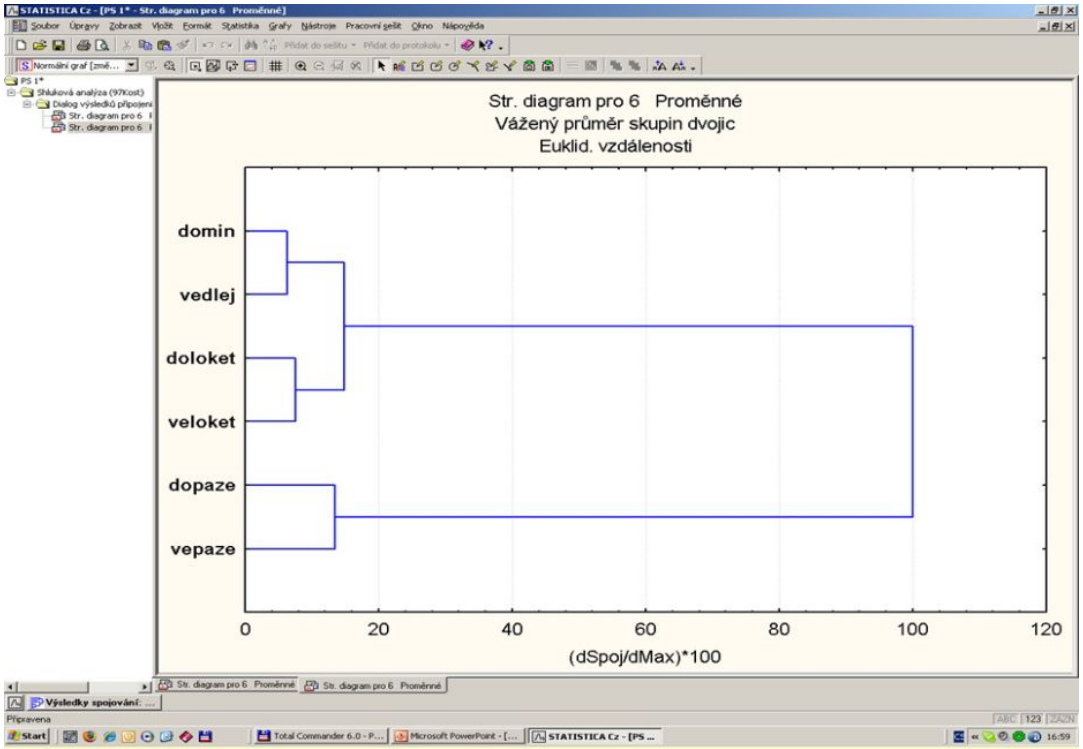
Dendrogram znaků matice dat *Kost* (STATISTICA).



Dendrogram 25 pacientů matice dat *Kost*, (STATISTICA).

○ ***Závěr:*** Pacienti byli roztříděni do třech shluků. Ostatní je třeba považovat za odlehlé.





PŘÍKLAD 9.13 *Klasifikace vlastností rozličných druhů kávy*

Byl získán výběr 43 vzorků kávy, pocházejících ze 30 zemí. U každého druhu kávy byly změřeny jeho chemické a fyzikální vlastnosti. Splňují data požadavky na homogenitu a je možné indikovat dvě či více rozličných kategorií? Vytvořte dendrogram klasifikovaných druhů kávy.

○ ***Data:*** Soubor dat *Kava* obsahuje 2 druhy kávy, Robusta a Arabica ve 43 vzorcích ze 30 zemí a popsaných 13 fyzikálně-chemickými znaky:

i značí index kávy,
Objekt značí původ kávy,
Voda značí obsah vody x_1,
Zrno značí hmotnost zrn x_2,
Extrakt značí extrakt x_3,
pH značí hodnotu pH x_4,
Acidita značí hodnotu volné acidity x_5,
Mineral značí obsah minerálů x_6,
Tuky značí obsah tuků x_7,
Kofein značí obsah kofeinu x_8,
Trinonelin značí obsah trinonelinu x_9,
Kchlorogen značí obsah kyseliny chlorogenikové x_{10},
Kneochlor značí obsah kyseliny neochlorogenikové x_{11},
Kisochlor značí obsah kyseliny isochlorogenikové x_{12},
Sumakys značí sumu kyselin chlorogenikových x_{13}.

i	*Objekt*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1.0	5.4	0.4	0.8	6.6
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5

○ ***Řešení:*** Graf komponentních vah znaků odhaluje především korelaci znaků. Je-li úhel mezi průvodiči dvou znaků malý, jsou dva znaky v silné korelaci.

První shluk obsahuje znaky *Voda, pH, Kchlorogen, Sumakysel, Mineral, Kofein, Trinonelin, Kneochlor, Kizochlor,* a *Tuky*.

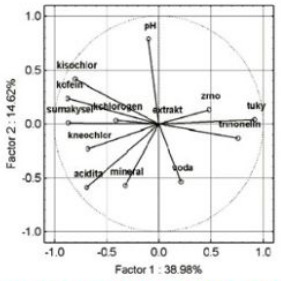
Druhý shluk obsahuje dva znaky, *Extrakt* a *Acidita.*

Vznik shluků druhů kávy lze sledovat na grafu komponentního skóre objektů a na dendrogramu objektů.

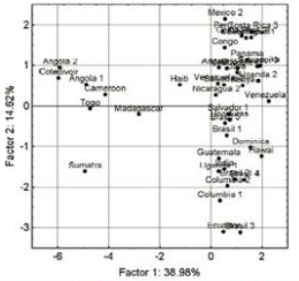
Graf komponentního skóre ukazuje, že 43 objektů čili druhů kávy v datovém souboru *Kava* nejsou dostatečně homogenní.

Objekty zde lze rozdělit do dvou shluků, v prvním vlevo je 7 objektů a ve druhém svislém shluku vpravo je zbývajících 36 objektů. Klasifikace do těchto shluků je především vlivem znaků *Tuky, Kofein, Trinonelin* a *Sumakys*.

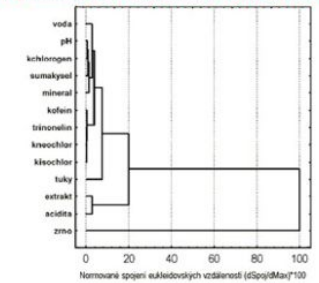
Na **dendrogramu objektů** při postupu zprava doleva je zřejmé, že druhy kávy lze rozdělit do dvou velkých shluků. Větší shluk nazvaný Arabica lze dále rozdělit na dva menší shluky Arabica A a Arabica B a jeden odlehlý objekt. Ve spodní části obrázku zůstává jeden větší shluk 13 druhů kávy, patřících zřejmě do druhu Robusta.



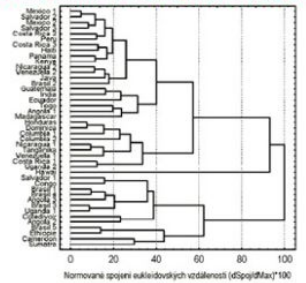
Graf komponentních vah znaků matice dat *Kava*, (STATISTICA).



Graf komponentního skóre objektů matice dat *Kava*

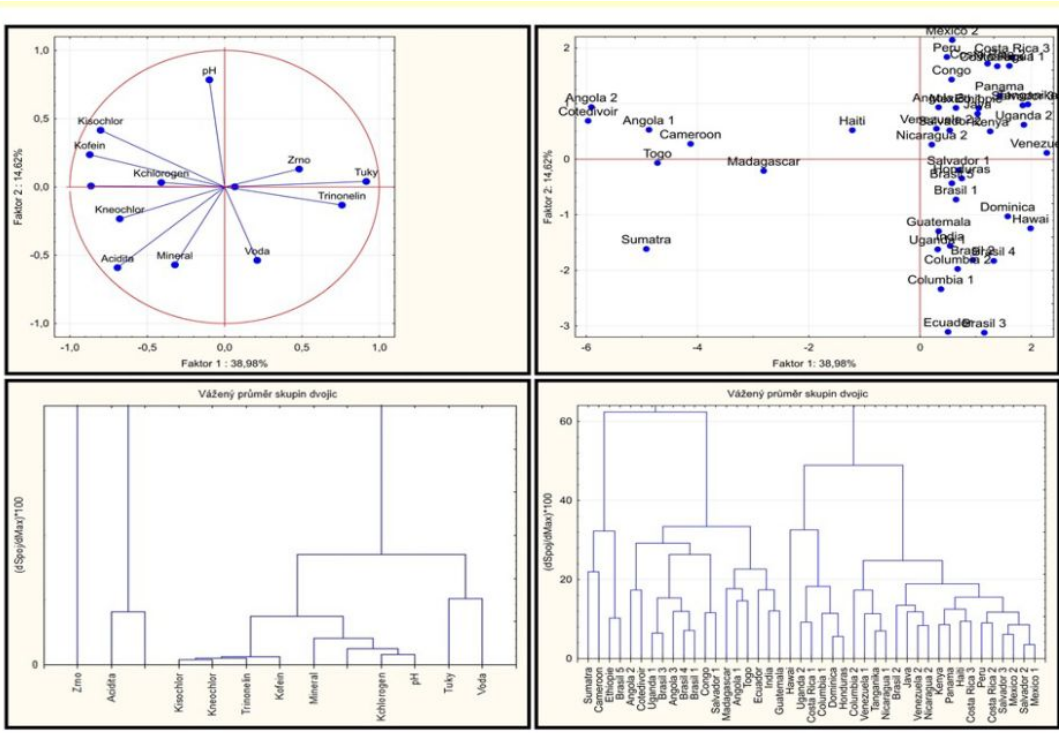


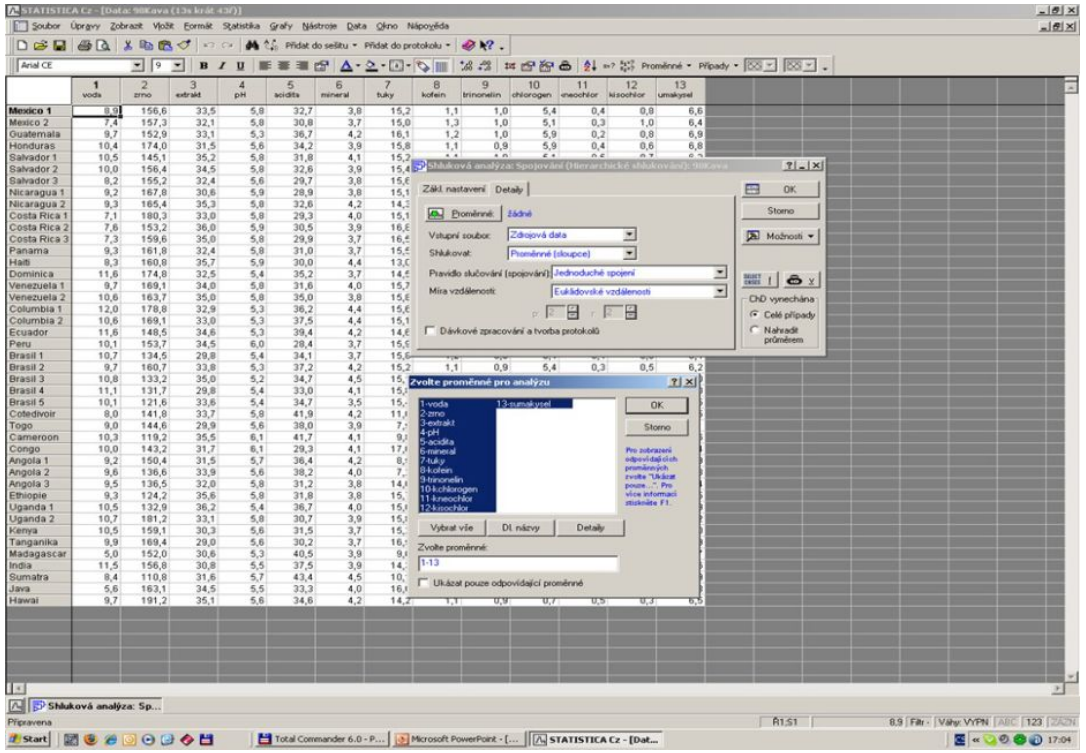
Dendrogram znaků matice dat *Kava* (STATISTICA).

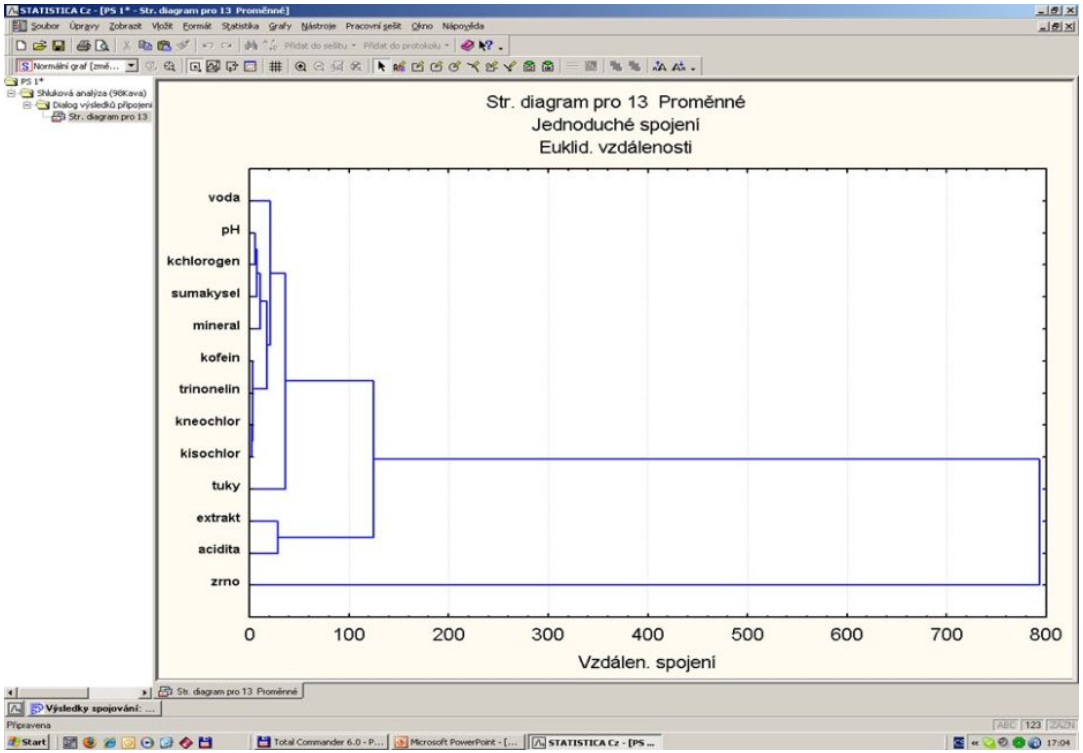


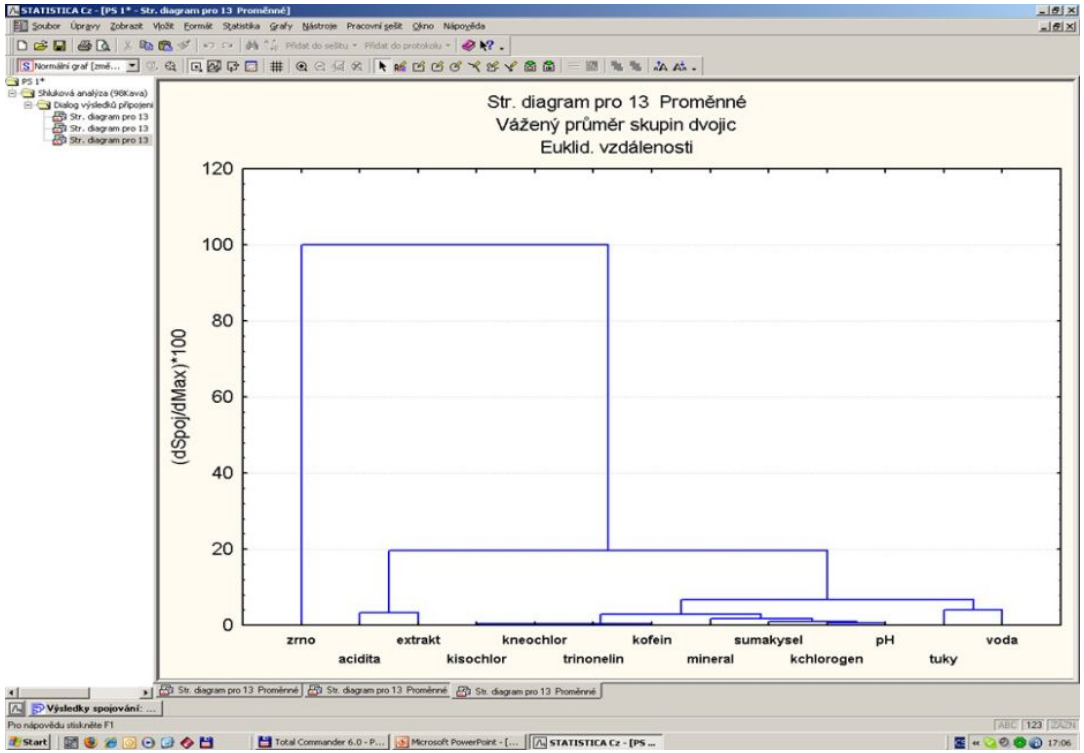
Dendrogram objektů matice dat *Kava*, (STATISTICA).

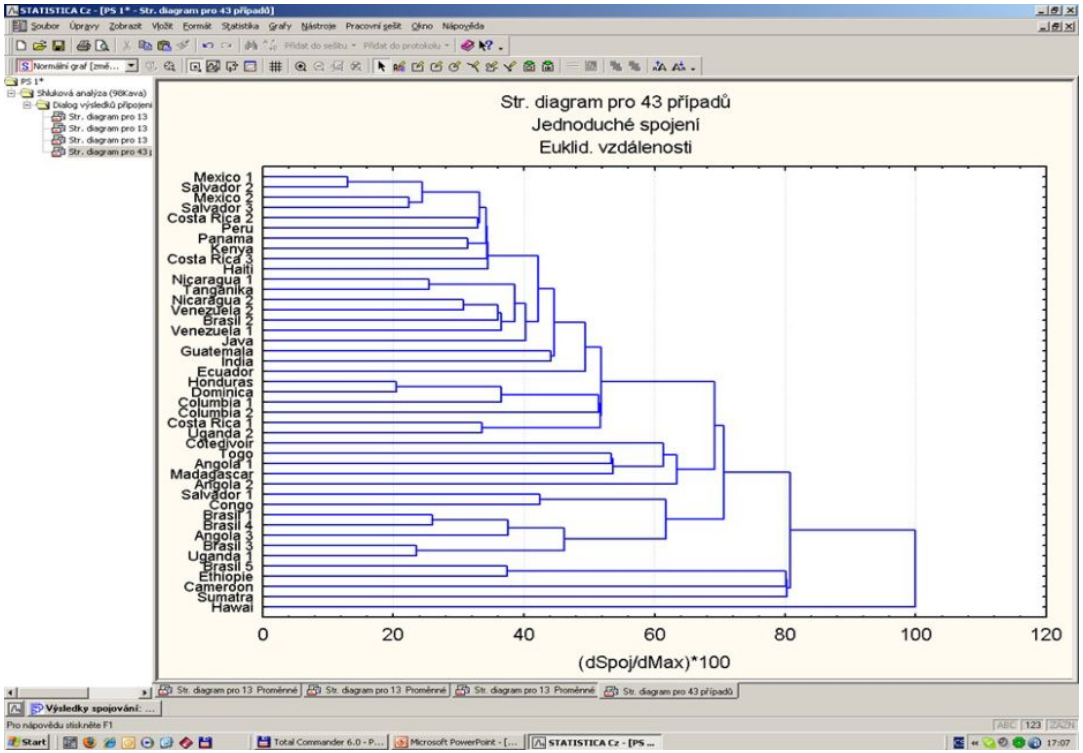
○ ***Závěr:*** Dendrogram znaků ukazuje shluky podobných vlastností kávy, zatímco dendrogram objektů klasifikuje podobné druhy kávy do shluků.

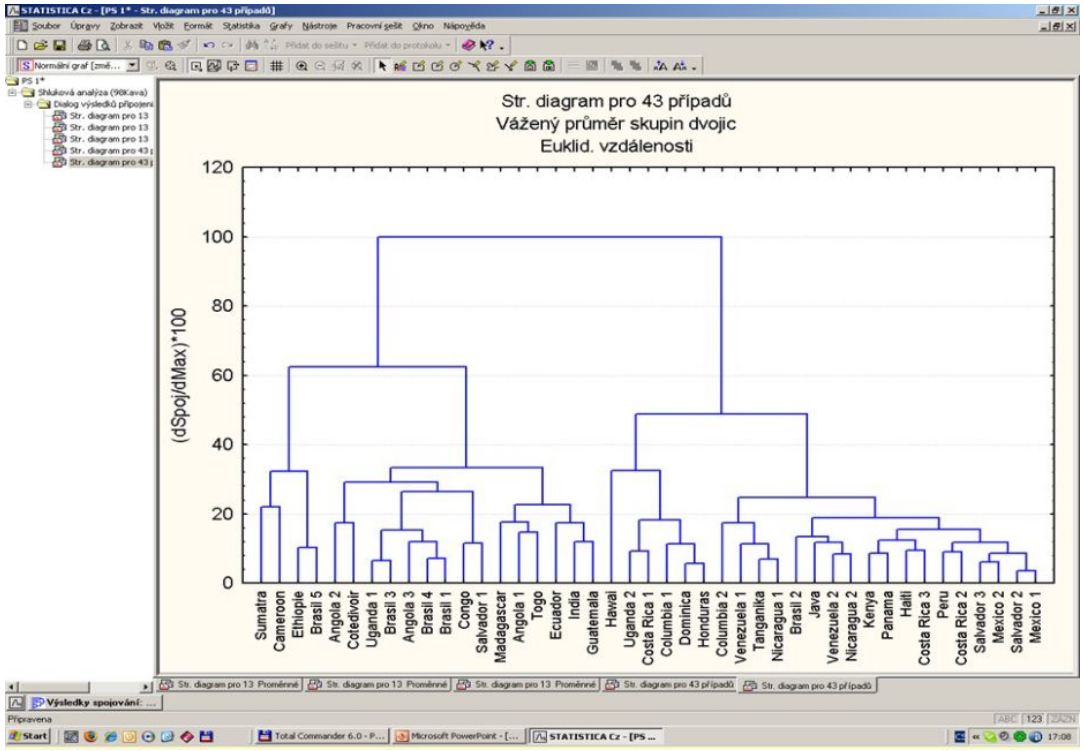


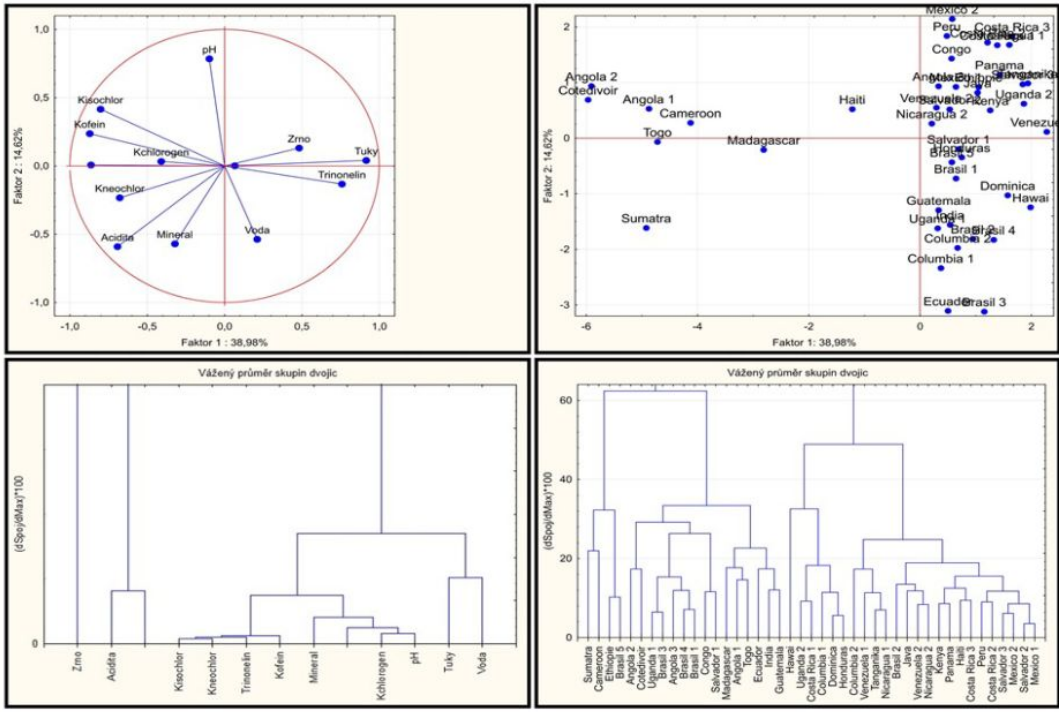



Str. diagram pro 13 Proměnné
Jednoduché spojení
Euklid. vzdálenosti
voda
pH
kchlorogen
sumakysel
mineral
kofein
trinonelin
kneochlor
kisochlor
tuky
extrakt
acidita
zrno
0
100
200
300
400
500
600
700
800
Vzdálen. spojení


Str. diagram pro 13 Proměnné
Vážený průměr skupin dvojic
Euklid. vzdálenosti
(dSpoj/dMax)*100
0
20
40
60
80
100
120
zrno
acidita
extrakt
kisochlor
kneochlor
trinonelin
kofein
mineral
sumakysel
kchlorogen
pH
tuky
voda


Str. diagram pro 43 případů
Jednoduché spojení
Euklid. vzdálenosti
Mexico 1
Salvador 2
Mexico 2
Salvador 3
Costa Rica 2
Peru
Panama
Kenya
Costa Rica 3
Haiti
Nicaragua 1
Tanganika
Nicaragua 2
Venezuela 2
Brasil 2
Venezuela 1
Java
Guatemala
India
Ecuador
Honduras
Dominica
Columbia 1
Columbia 2
Costa Rica 1
Uganda 2
Coledivoir
Togo
Angola 1
Madagascar
Angola 2
Salvador 1
Congo
Brasil 1
Brasil 4
Angola 3
Brasil 3
Uganda 1
Brasil 5
Ethiopie
Camerdon
Sumatra
Hawai
0
20
40
60
80
100
120
(dSpoj/dMax)*100


Str. diagram pro 43 případů
Vážený průměr skupin dvojic
Euklid. vzdálenosti
(dSpoj/dMax)*100
0
20
40
60
80
100
120
Sumatra
Cameroon
Ethiopie
Brasil 5
Angola 2
Coledivoir
Uganda 1
Brasil 3
Angola 3
Brasil 4
Brasil 1
Congo
Salvador 1
Madagascar
Angola 1
Togo
Ecuador
India
Guatemala
Hawai
Uganda 2
Costa Rica 1
Columbia 1
Dominica
Honduras
Columbia 2
Venezuela 1
Tanganika
Nicaragua 1
Brasil 2
Java
Venezuela 2
Nicaragua 2
Kenya
Panama
Haiti
Costa Rica 3
Peru
Costa Rica 2
Salvador 3
Mexico 2
Salvador 2
Mexico 1



PŘÍKLAD 9.14 *Klasifikace vzorků italských vín*

Pro 90 vzorků italských vín bylo naměřeno 8 fyzikálně-chemických vlastností. Ve vínech jsou obsaženy tři kultury, a to Nebbiolo ve víně Barolo, Grignolino a Barbera ve vínech stejného jména, a to každá ve 30 vzorcích. Kolik faktorů rozliší tři kategorie vín? Do kolika shluků lze vína roztřídit? Souvisí počet shluků se zadanými druhy vín?

○ ***Data:*** Soubor dat ***Vina*** je popsán:

i značí index vzorku vína,
Objekt značí jméno vzorku vína,
Kateg značí kategorie vzorku vína a 90 druhů vín v řádcích se týká tří kategorií 1. Barolo, 2. Grignolino a 3. Barbera, popsaných 8 následujícími vlastnostmi čili znaky ve sloupcích:
Alkohol značí obsah alkoholu x_1,
Necuk značí necukerný extrakt x_2,
Fosfaty značí obsah fosfátů x_3,
Fenoly značí obsah celkových fenolů x_4,
Flavan značí obsah flavanoidů x_5,
PomerA1 značí naměřený poměr absorbancí při 280 a 315 nm pro naředěné víno x_6,
PomerA2 značí naměřený poměr absorbancí při 280 a 315 nm pro určení flavanoidů x_7,
Prolin značí obsah prolinu x_8.

i	*Objekt*	*Kateg*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	Olo0171	1	14.23	24.82	320	2.80	3.06	3.92	4.77	1065
..	..	..	..	..	..	..	..	..	..	..
90	Era2878	3	13.17	23.45	534	1.65	0.68	1.62	2.05	840

○ ***Řešení:*** **Graf komponentních vah znaků** odhaluje především korelaci znaků. Blízké průvodiče znaků s malým úhlem indikují silnou pozitivní korelaci znaků.

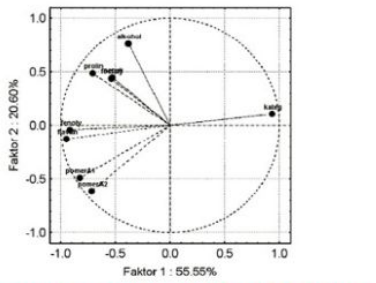
Dendrogram znaků ukazuje na první shluk podobných znaků *Fenoly, Flavan, PomerA1, PomerA2, Alkohol, Necuk* a také *Kateg*.

K tomuto shluku se pojí již podstatně méně podobný znak *Fosfaty*.

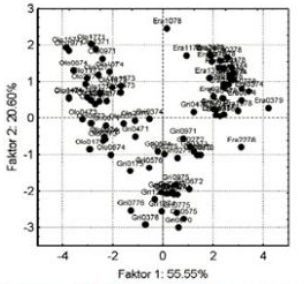
Znak *prolin* je zcela nepodobný ostatním a v dendrogramu je indikován jako odlehlý znak.

Graf komponentního skóre objektů vykazuje tři větší shluky vín ve shodě s jejich kategoriemi *Barolo* ve zkratce *Olo,* Barbera ve zkratce *Era* a konečně *Grignolino* ve zkratce *Gri.*

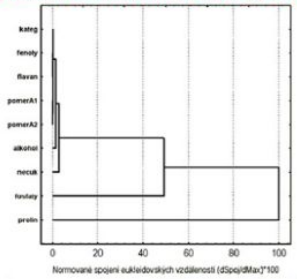
Dendrogram objektů rovněž ukazuje na tři shluky, zhora první shluk *Olo*, uprostřed grafu shluk *Era* a v dolní části grafu pak shluk *Gri*.



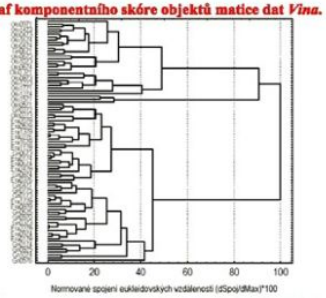
Graf komponentních vah znaků matice dat *Vina*, (STATISTICA).



Graf komponentního skóre objektů matice dat *Vina*.

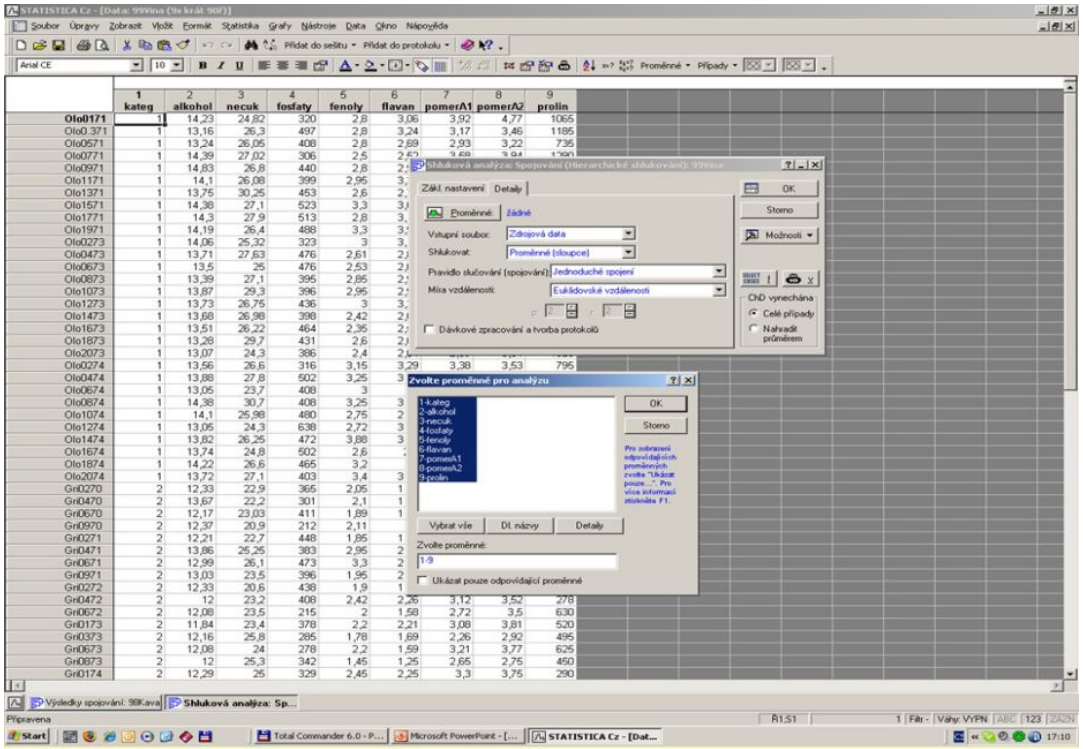


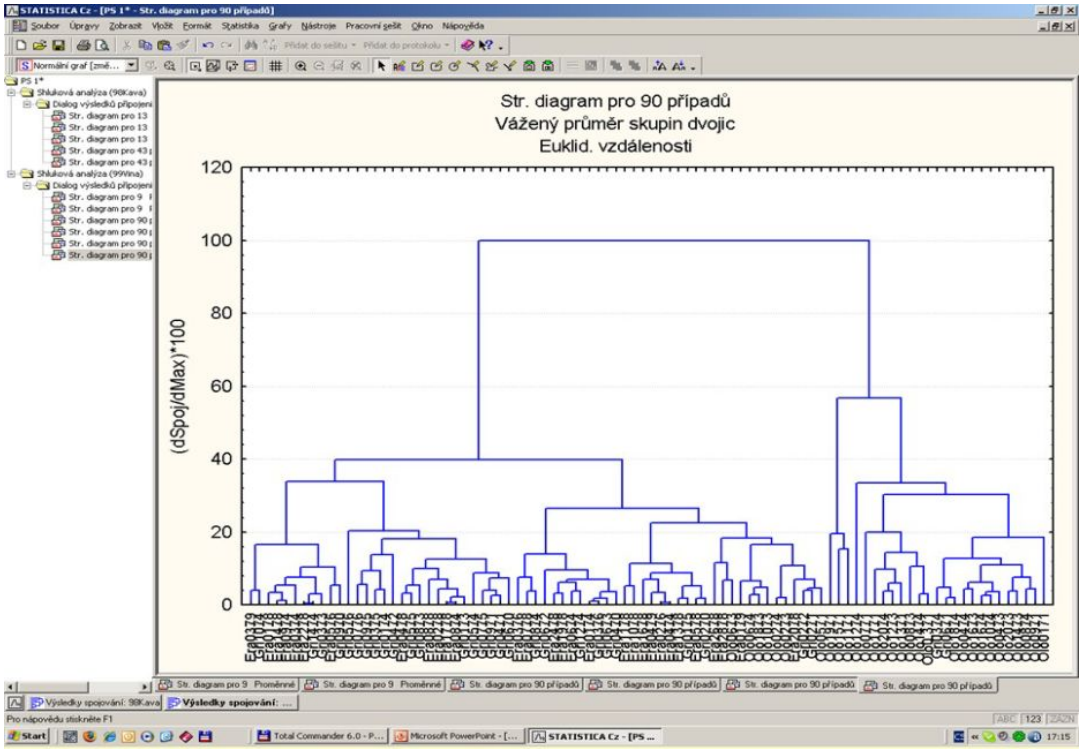
Dendrogram znaků matice dat *Vina* (STATISTICA).



Dendrogram objektů matice dat *Vina*, (STATISTICA).

○ ***Závěr:*** Graf komponentního skóre objektů a dendrogram objektů shodně vykazují tři větší shluky vín ve shodě s jejich kategoriemi ***Olo, Era*** a ***Gri***.





PŘÍKLAD 9.15 *Hledání podobnosti vlastností křupavých lupínků od různých výrobců*

Tři americké firmy General Mills (G), Kellogg (K) a Quaker (Q) produkují křupavé obilné lupínky a bylo sledováno 10 znaků. Byla vyšetřována struktura a vzájemné vazby mezi sledovanými znaky jednotlivých produktů, ale i mezi objekty. Které objekty jsou si velice podobné?

Data: Datová matice *Krupky* obsahuje 55 dodavatelů a vyšetřováno 10 znaků:

Objekt značí index obilných lupínků x_1,
i značí jednoho ze tří výrobců G, K či Q x_2,
Cal značí kalorickou hodnotu [cal] x_3,
Bilkov značí obsah bílkovin x_4,
Tuky značí obsah tuků x_5,
Na značí obsah sodných iontů x_6,
Vlakn značí obsah vlákniny x_7,
Uhlovod značí obsah uhlovodíků x_8,
Cukr značí obsah cukru x_9,
K značí obsah draselných iontů x_{10},
Skupina značí zařazení do skupiny x_{11}.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	ACCheerios	G	110	2	2	180	1.5	10.5	10	70	1
...	...	...	...	...	...	...	...	...	...	...	..
55	QuakerOatmeal	Q	100	5	2	0	2.7	1	1	110	3

○ ***Řešení:*** Korelaci znaků indikuje graf komponentních vah znaků.

Tři znaky *Na, Cal, Cukr* jsou v silné korelaci, protože jsou v grafu blízko sebe a úhel mezi jejich průvodiči je velice malý.

Druhý shluk obsahuje čtyři znaky *Tuky, K, Vlakn, Bilkov*, které jsou vzájemně rovněž silně korelovány.

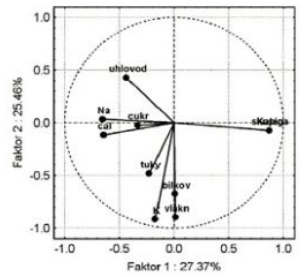
***Skupina* a *Kategorie* korelují**, protože označují stejnou věc.

Uhlovod je vybočující znak, který slabě či vůbec nekoreluje s ostatními znaky.

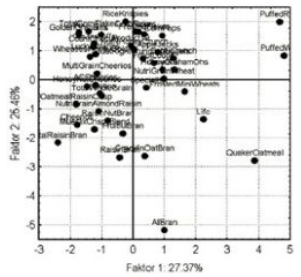
Dendrogram znaků ukazuje dva shluky a dva zcela odlehlé znaky:
První shluk obsahuje 6 vzájemně velice podobných znaků *Bilkov, Vlakn, Tuky, Skupina, Cukr* a *Uhlovod*.
Druhý shluk obsahuje 2 znaky *Kateg* a *Cal*. K nim se připojuje osamocený znak *K*.

Naprosto nepodobný znak vůči všem ostatním znakům je *Na*.

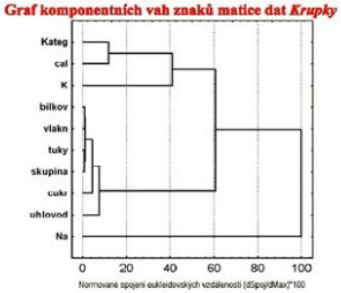
Graf komponentního skóre objektů naznačuje několik shluků objektů, které jsou v souladu se shluky určenými na základě eukleidovské vzdálenosti v dendrogramu. Zcela nepodobný objekt se všemi ostatními se jeví *AllBran*. Také další tři objekty se jeví silně odlišné od ostatních.



Graf komponentních vah znaků matice dat *Krupky*



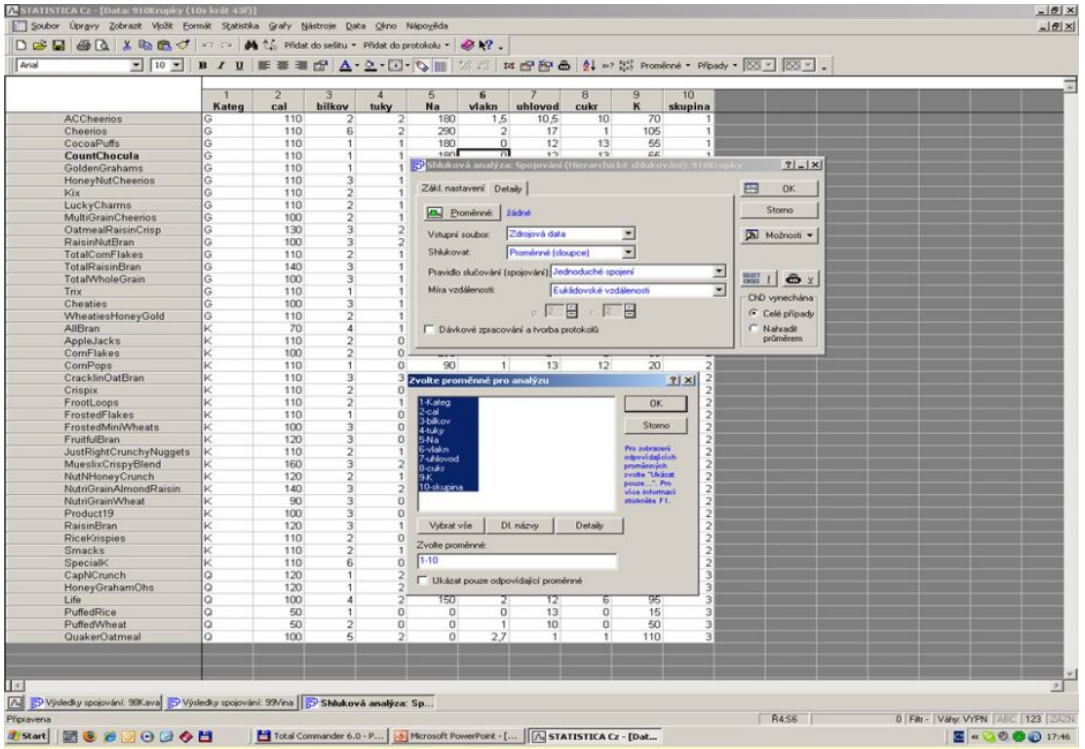
Graf komponentního skóre objektů matice dat *Krupky*

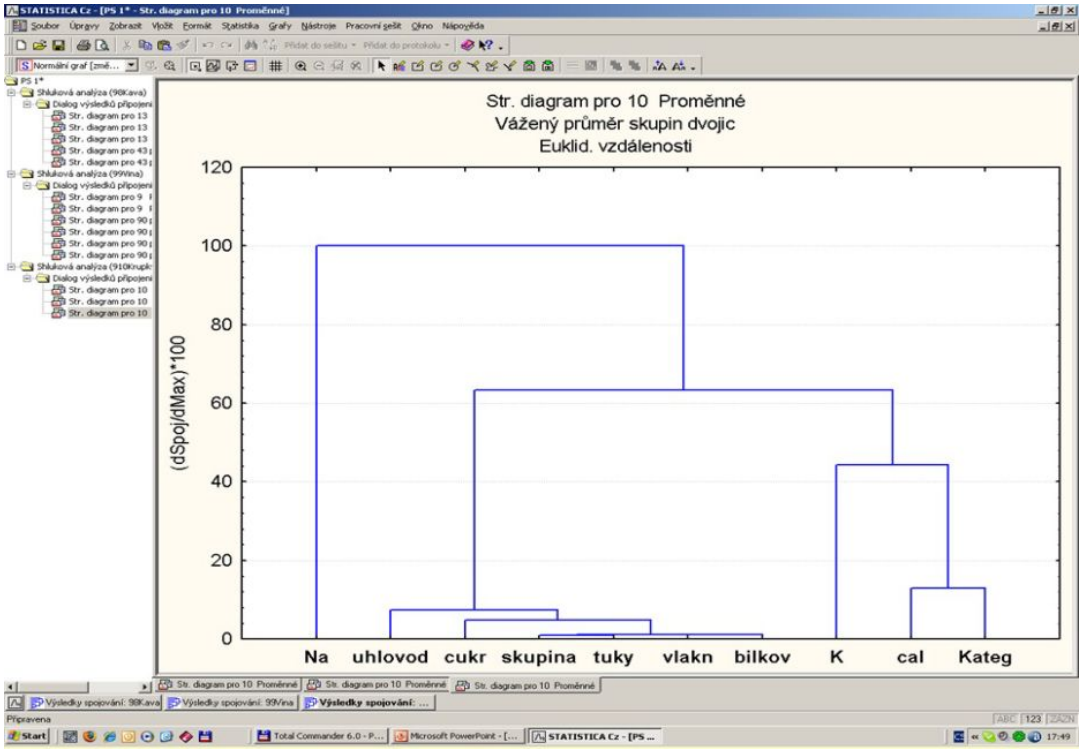


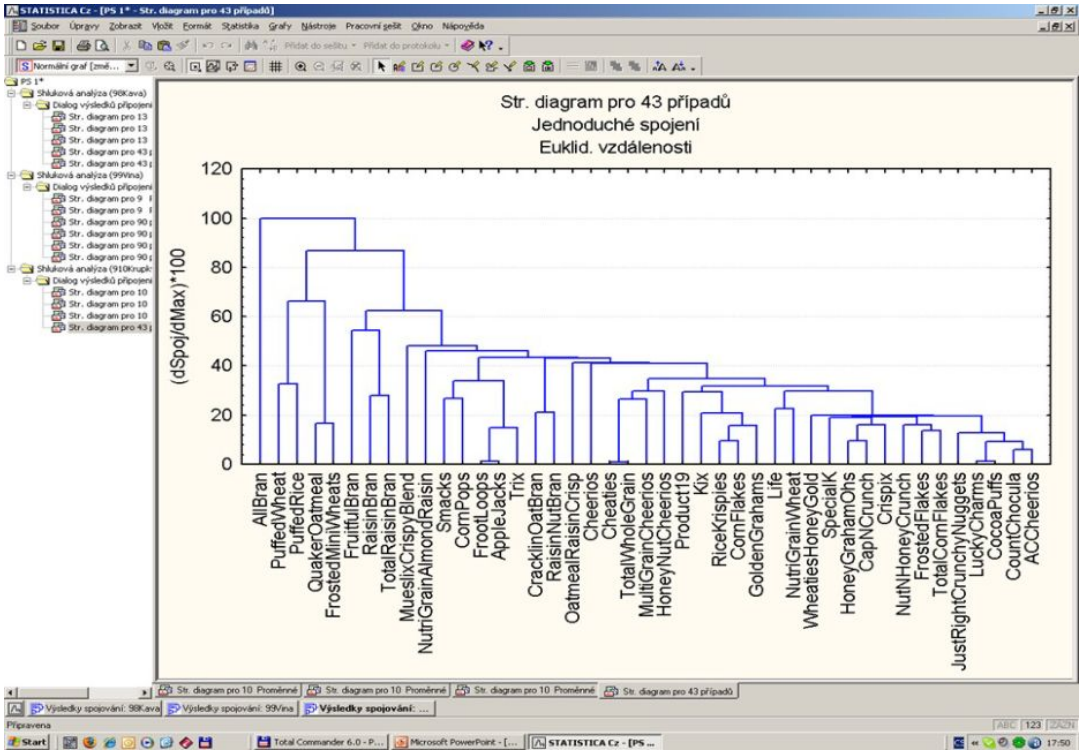
Dendrogram znaků matice dat *Krupky*, (STATISTICA).

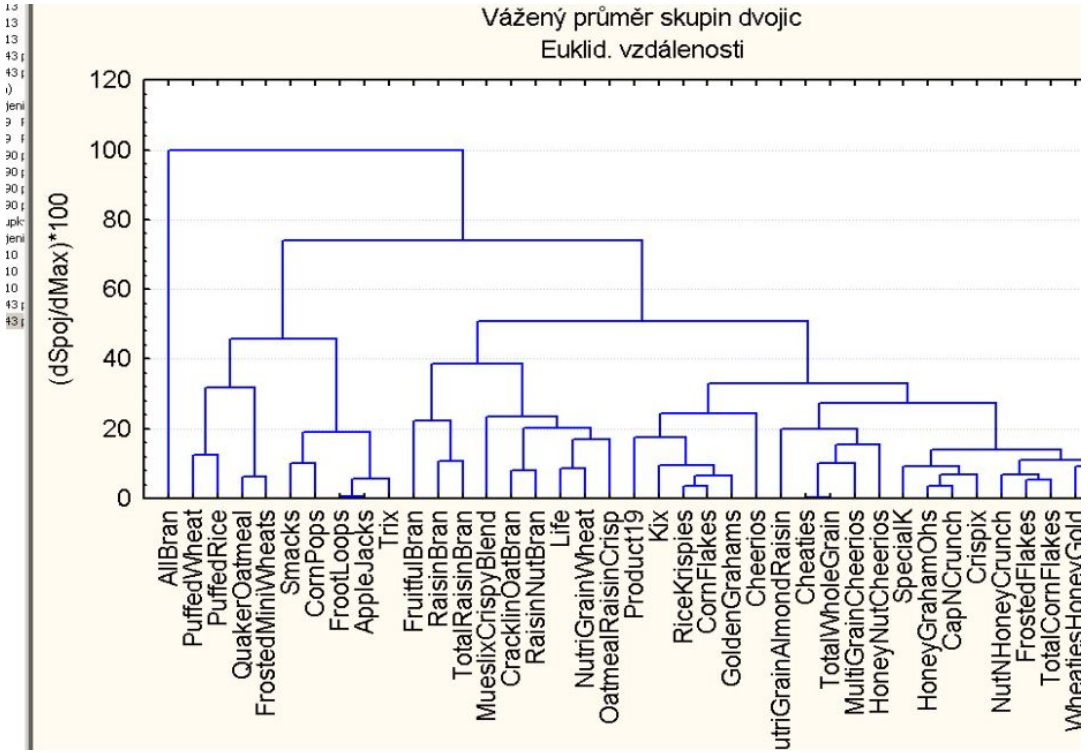
Dendrogram objektů matice dat *Krupky*, (STATISTICA).

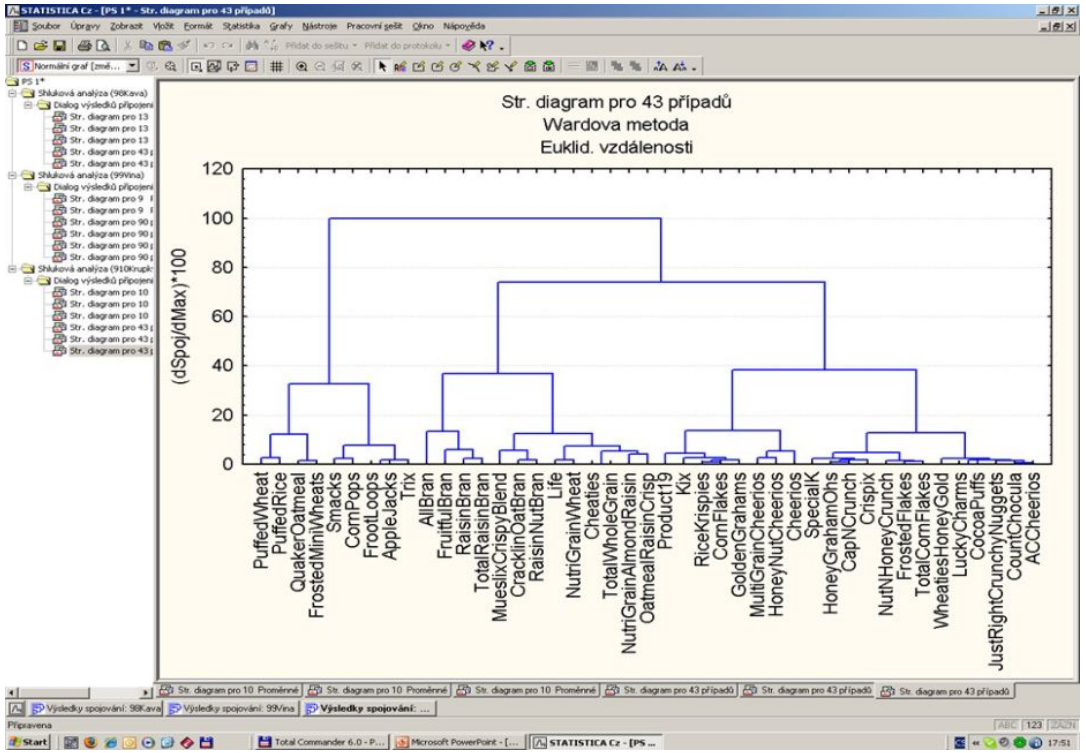
○ ***Závěr:*** Shlukováním metodou skupinového průměru se podařilo najít několik druhů křupavých lupínků, které jsou zcela nepodobné ostatním.








Vážený průměr skupin dvojic
Euklid. vzdálenosti
(dSpoj/dMax)*100
120
100
80
60
40
20
0
AllBran
PuffedWheat
PuffedRice
QuakerOatmeal
FrostedMiniWheats
Smacks
CornPops
FrootLoops
AppleJacks
Trix
FruitfulBran
RaisinBran
TotalRaisinBran
MueslixCrispyBlend
CracklinOatBran
RaisinNutBran
Life
NutriGrainWheat
OatmealRaisinCrisp
Product19
Kix
RiceKrispies
CornFlakes
GoldenGrahams
Cheerios
utriGrainAlmondRaisin
Cheaties
TotalWholeGrain
MultiGrainCheerios
HoneyNutCheerios
SpecialK
HoneyGrahamOhs
CapNCrunch
Crispix
NutNHoneyCrunch
FrostedFlakes
TotalCornFlakes
WheatiesHoneyGold


Str. diagram pro 43 případů
Wardova metoda
Euklid. vzdálenosti
(dSpoj/dMax)*100
120
100
80
60
40
20
0
PuffedWheat
PuffedRice
QuakerOatmeal
FrostedMiniWheats
Smacks
CornPops
FrootLoops
AppleJacks
Trix
AllBran
FruitfulBran
RaisinBran
TotalRaisinBran
MueslixCrispyBlend
CracklinOatBran
RaisinNutBran
Life
NutriGrainWheat
Cheaties
TotalWholeGrain
NutriGrainAlmondRaisin
OatmealRaisinCrisp
Product19
Kix
RiceKrispies
CornFlakes
GoldenGrahams
MultiGrainCheerios
HoneyNutCheerios
Cheerios
SpecialK
HoneyGrahamOhs
CapNCrunch
Crispix
NutNHoneyCrunch
FrostedFlakes
TotalCornFlakes
WheatiesHoneyGold
LuckyCharms
CocoaPuffs
JustRightCrunchyNuggets
CountChocula
ACCheerios