NCSS2007
ve vícerozměrné statistické analýze

Analýza shluků

CLU

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.

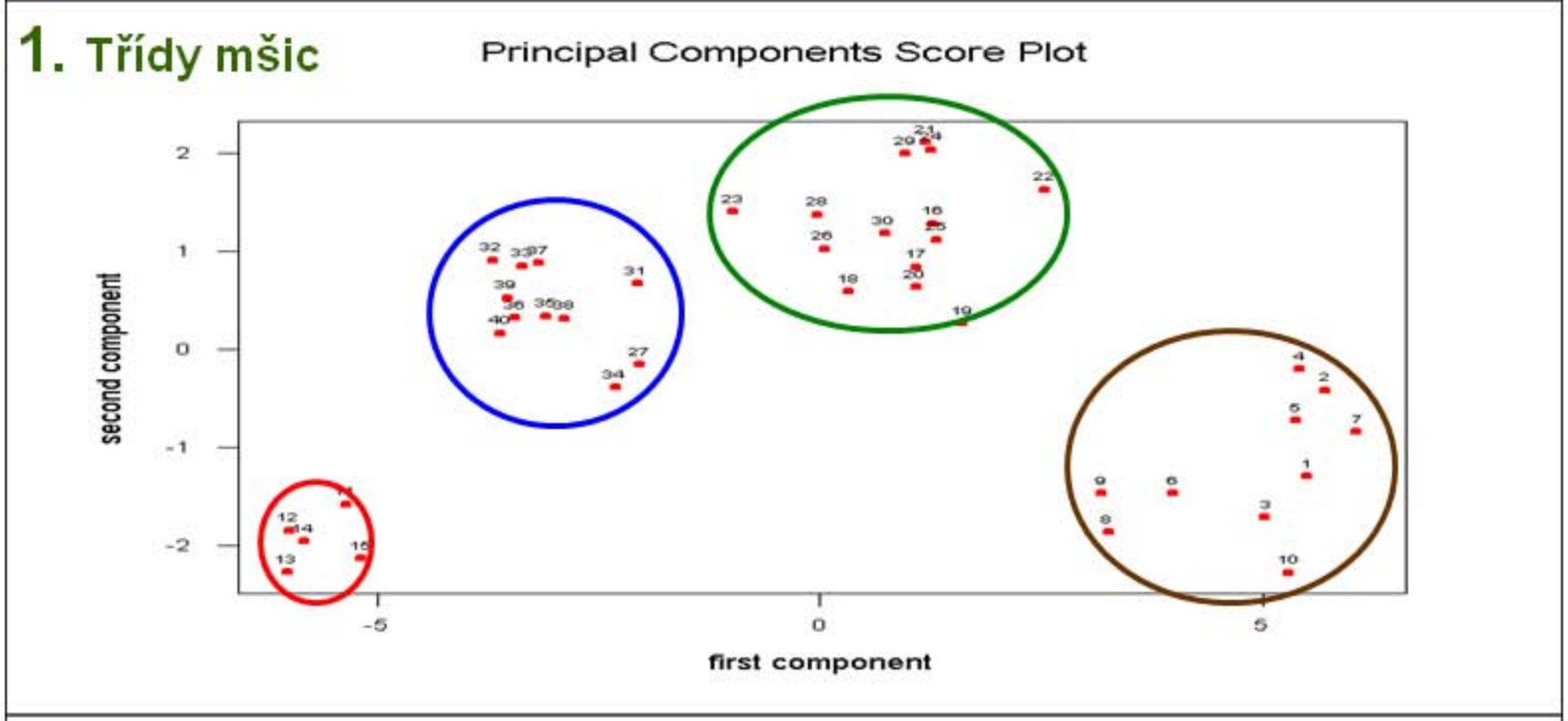
Identifikace vztahu: dle struktury shluků je snadnější odhalit vztahy mezi objekty. Shluky mohou být předmětem dalšího kvalitativního uvažování.

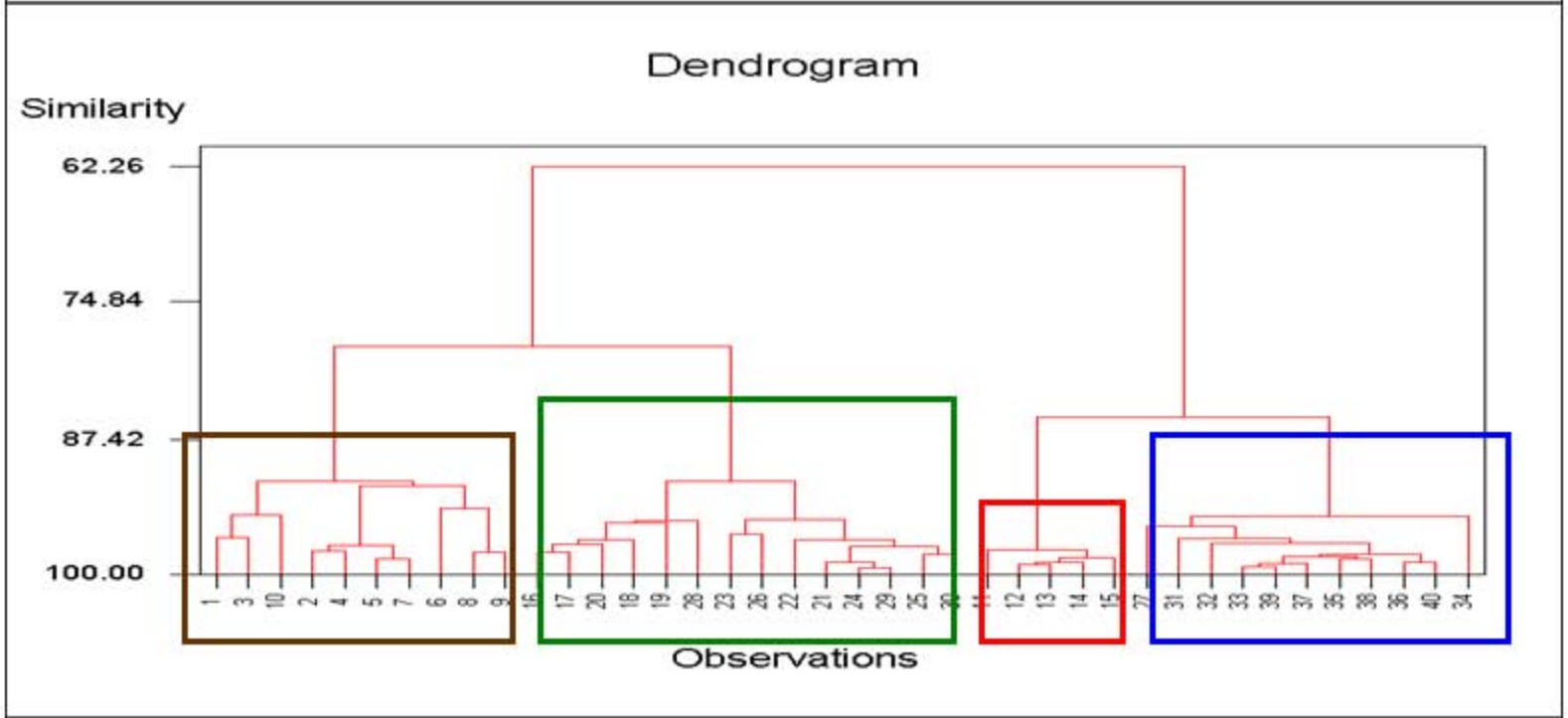
Úloha 1. Klasifikace polétavých mšic (Kompendium B404)

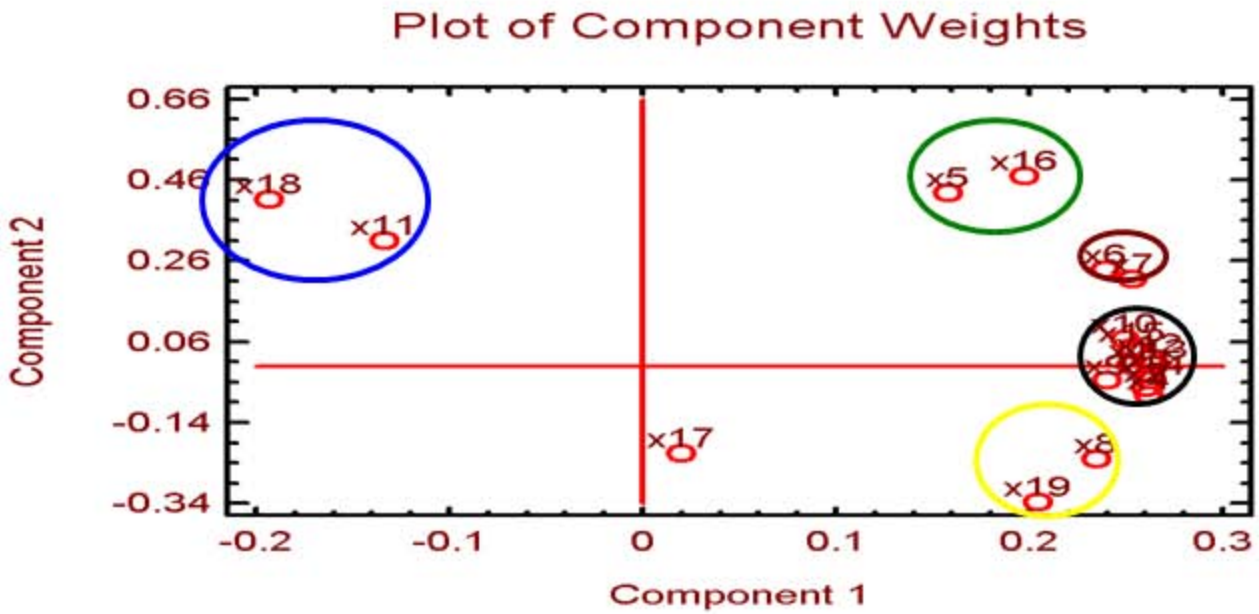
Jeffers (1967)25 studoval 40 jedinců polétavých mšic *(Alate adelges)* : 19 ukazatelů k rozlišení druhů, 14 znaků délky a šířky, 4 znaky se týkají počtu a 1 binární vyjadřuje přítomnost či absenci: x1 délka těla, x2 šířka těla, x3 délka předního křídla, x4 délka zadního křídla, x5 počet průduchů, x6 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

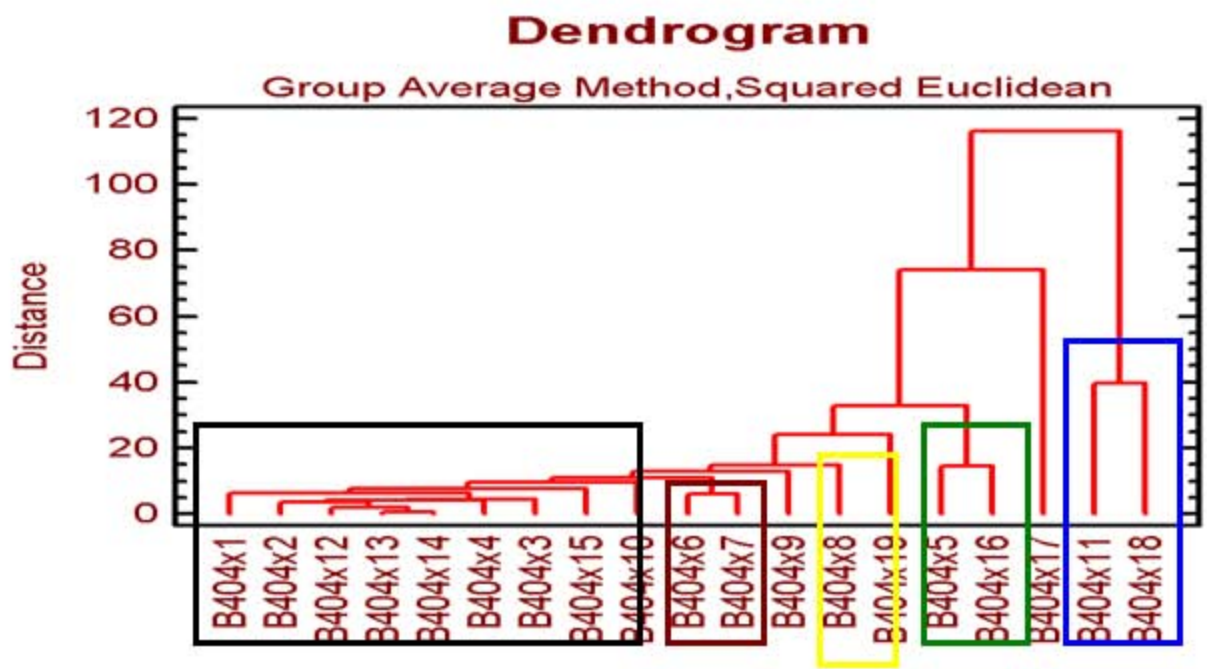
x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19
21.2	11	7.5	4.8	5	2	2	2.8	2.8	3.3	3	4.4	4.5	3.6	7	4	8	0	3
20.2	10	7.5	5	5	2.3	2.1	3	3	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10	7	4.6	5	1.9	2.1	3	2.5	3.3	1	4.2	4.4	3.3	7	4	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	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	6.6	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






Plot of Component Weights
Component 2
Component 1
0.66
0.46
0.26
0.06
-0.14
-0.34
-0.2
-0.1
0
0.1
0.2
0.3
x18
x11
x5
x16
x6
x7
x17
x8
x19


Dendrogram
Group Average Method,Squared Euclidean
Distance
120
100
80
60
40
20
0
B404x1
B404x2
B404x12
B404x13
B404x14
B404x4
B404x3
B404x15
B404x10
B404x6
B404x7
B404x9
B404x8
B404x19
B404x5
B404x16
B404x17
B404x11
B404x18

2. krok: Formulace úlohy analýzy shluků

Vyhodnotí všechny možné kombinace shluků,
u 25 objektů a 5 shluků existuje 2.4×10^{15} možných shluků.

Uživatel musí určit jediné správné řešení.

Návrh modelu shluků a použité techniky má větší důležitost
než u ostatních vícerozměrných technik.

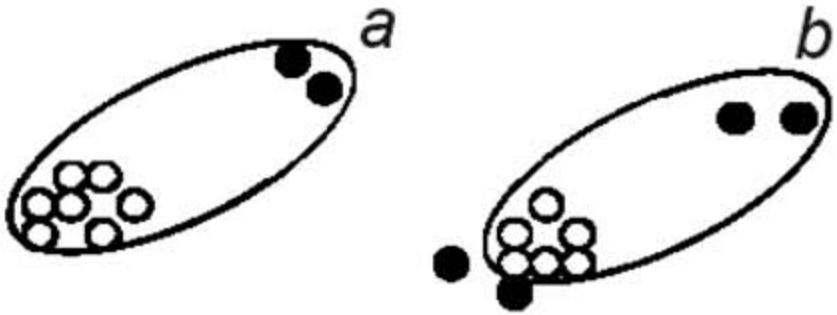
2.1 Odhalení odlehlých objektů, outlierů

Outliery představují
(1) odchýlené objekty, které nejsou představiteli populace,
(2) chybný výběr objektu z dané populace.

Outliery zbortí
(1) strukturu dat,
(2) nalezené shluky nebudou představovat skutečnou strukturu objektů dané populace.

Nalezení outlierů
profilovým diagramem proměnných.

Outliery z dat odstranit
někdy se zbortí aktuální struktura objektů.



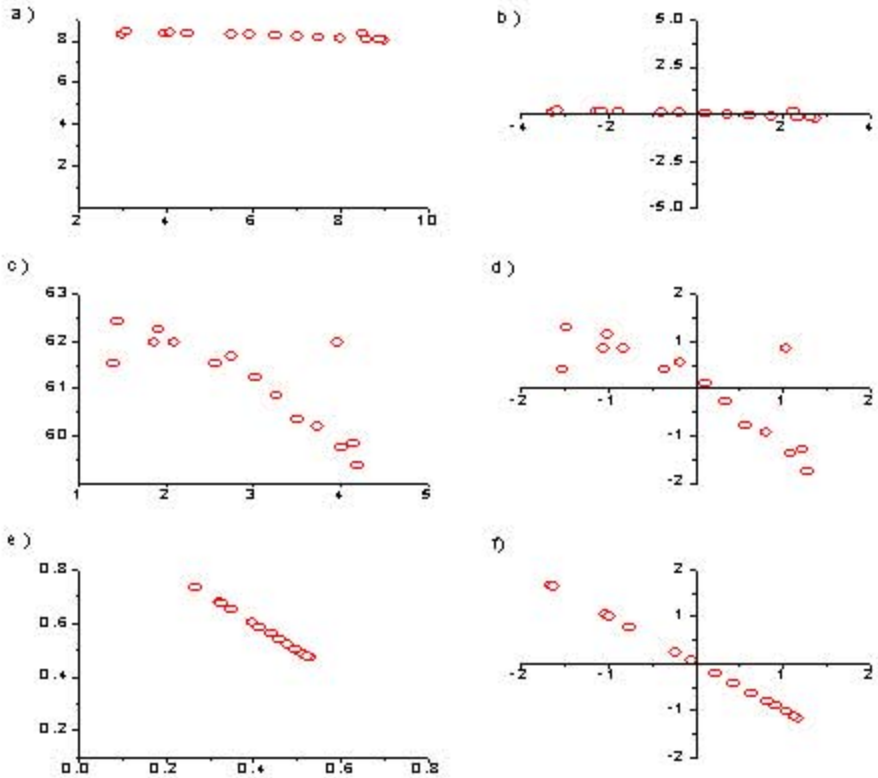
2.2 Standardizace dat

Aplikuje se před analýzou, míry vzdálenosti jsou citlivé na rozličné stupnice nebo na lišící se numerické velikosti proměnných.

Pravidlo: proměnné s větší proměnlivostí (směrodatnou odchylkou) mají větší vliv na míru podobnosti.

Standardizování proměnných

Standardizace je transformace proměnné do svého Z-skóre: (odečtením sloupcového průměru od každé hodnoty ve sloupci a výsledek se podělí sloupcovou směrodatnou odchylkou): Průměr standardizovaných dat je 0 se směrodatnou odchylkou 1.

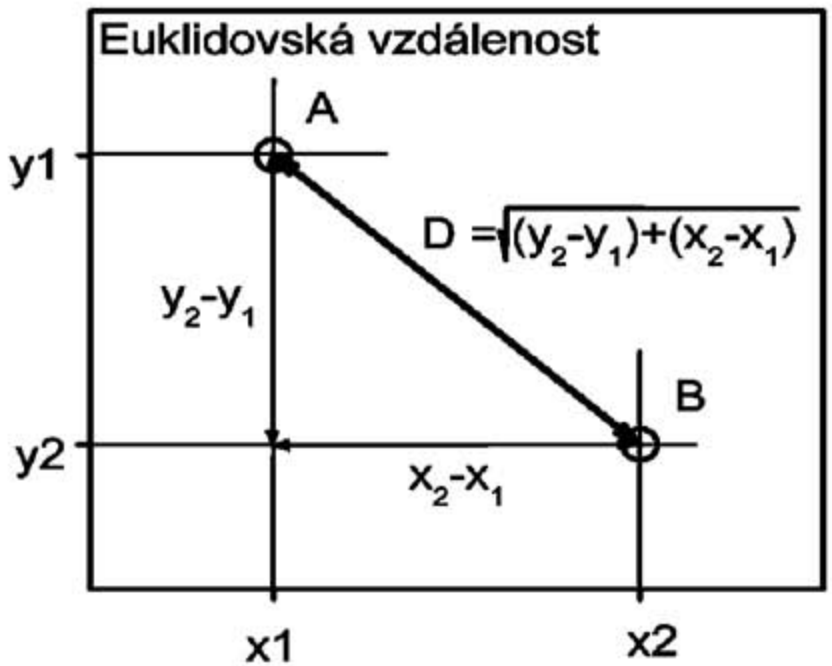


Efekt škálovacích technik:
(a) Originální data,
(b) sloupcové centrování,
(c) sloupcové standardizování,
(d) autoškálování,
(e) profily,
(f) autoškálované profily

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

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

3. krok: Předpoklady analýzy shluků

Analýza shluků není charakteru statistického testování.
Objektivní kvantifikace strukturních vlastností souboru objektů.
Nemá požadavky normality, linearity, homoskedasticity.

Existují pouze dva kritické předpoklady:

Reprezentativnost vzorku

Výběr objektů a odvozené shluky představují strukturu populace.
Zvolený výběr dat musí být opravdovým představitelem populace.
Odlehlé objekty zdůrazní divergentní shluky, které zanesou vychýlení do odhadu struktury objektů.
Výběr musí být dostatečně reprezentativní a výsledky zobecnitelné na celou populaci.

Vliv multikolinearity

Multikolineární proměnné jsou implicitně váženy intenzivněji.
Vyšetřit proměnné na přítomnost multikolinearity:

(1) Je třeba zredukovat počet proměnných
(2) Použít Mahalanobisovu vzdálenost.

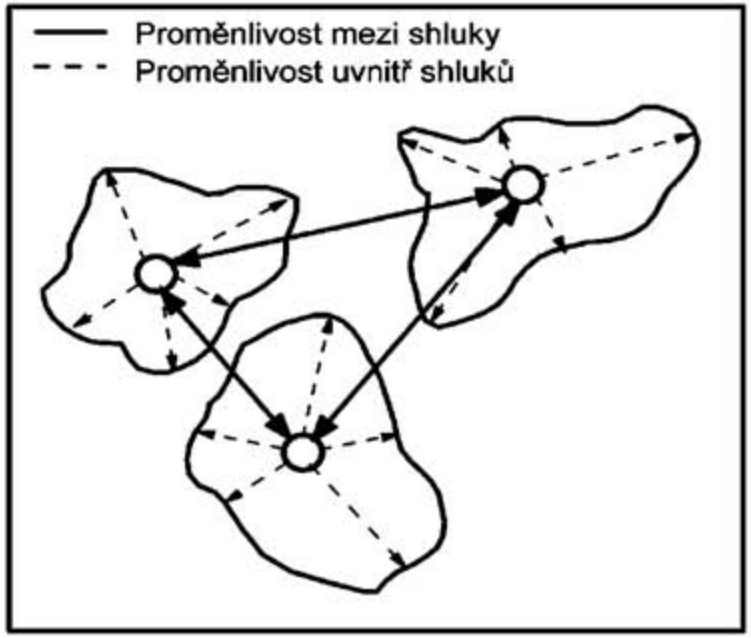
4. krok: Výstavba dendrogramu shluků

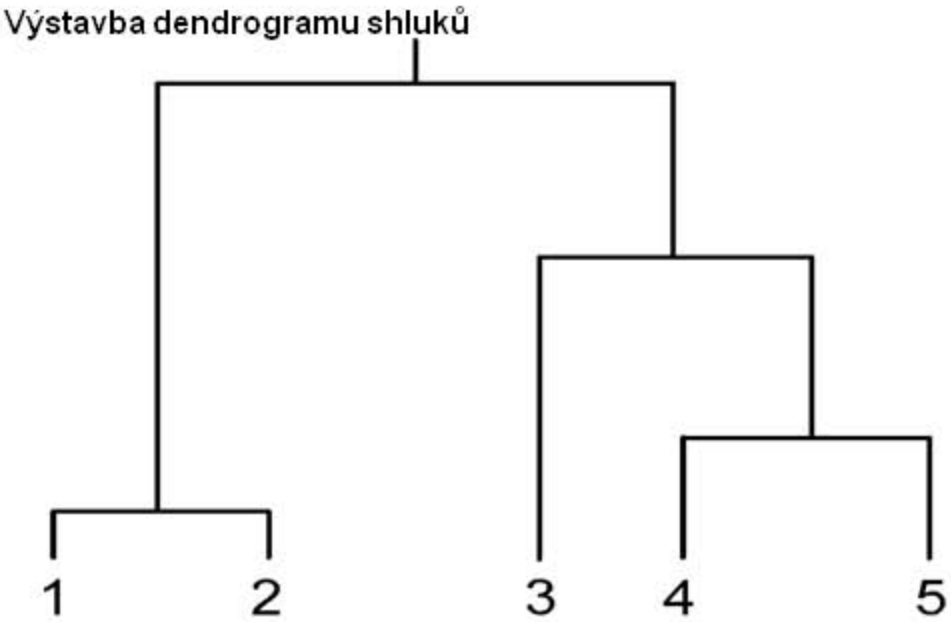
Vedle algoritmu je třeba vybrat i vhodný postup.

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

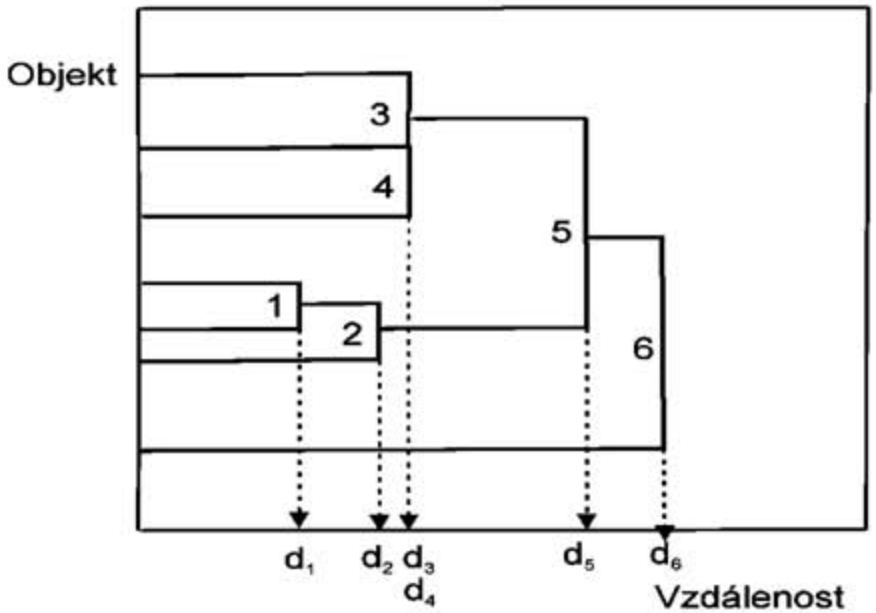
konstrukce stromovité struktury, dendrogramu

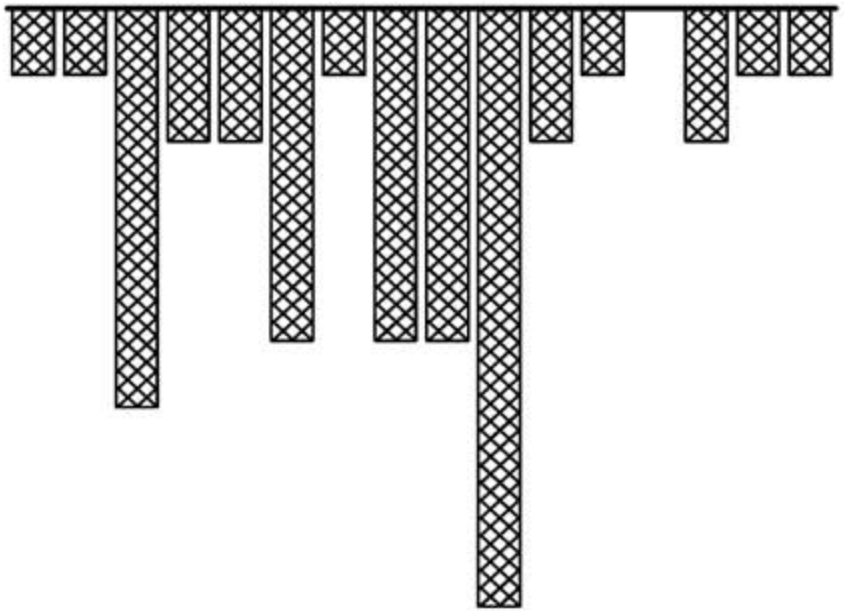
Způsoby hierarchického shlukování: aglomerační a divizní,

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.

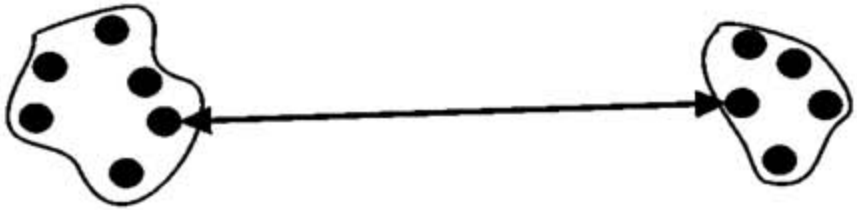
Metoda nejvzdálenějšího souseda: je postavena nikoliv na minimální ale na maximální vzdálenosti.

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.

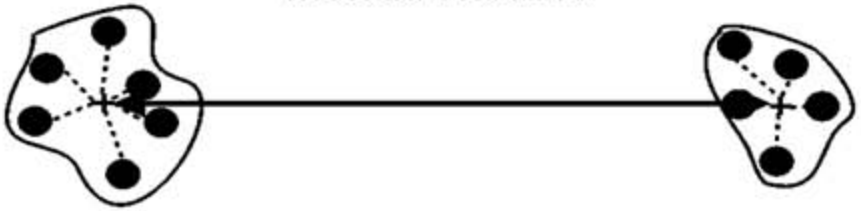
Metoda nejblizšího souseda



Metoda nevzdálenějšího souseda

Metoda průměrné vzdálenosti

Metoda Wardova



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

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

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)

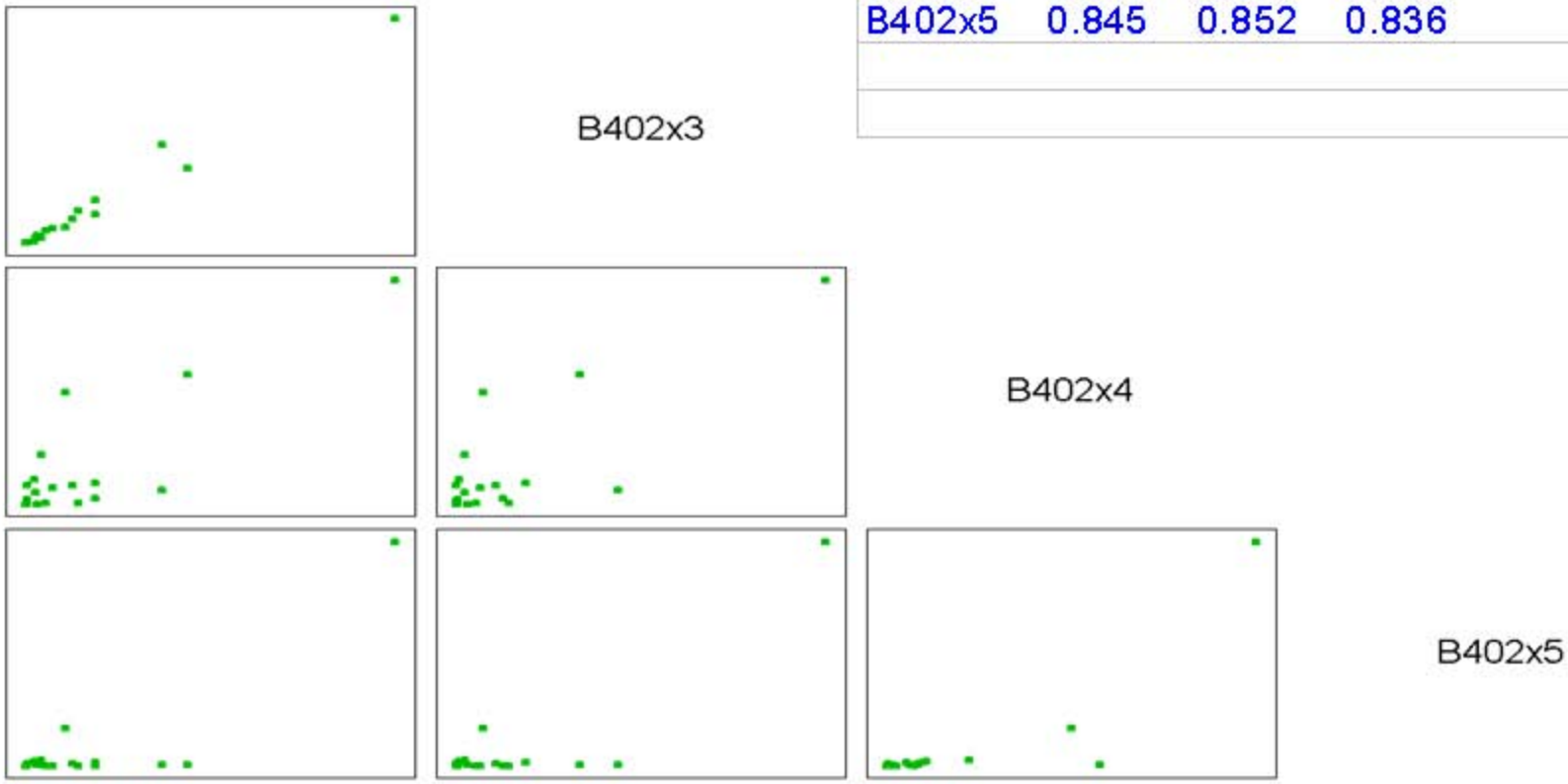
Liší se v účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění. Provedeme klasifikaci neuroleptik do shluků podobných účinků s ohledem na 4 znaky.

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:

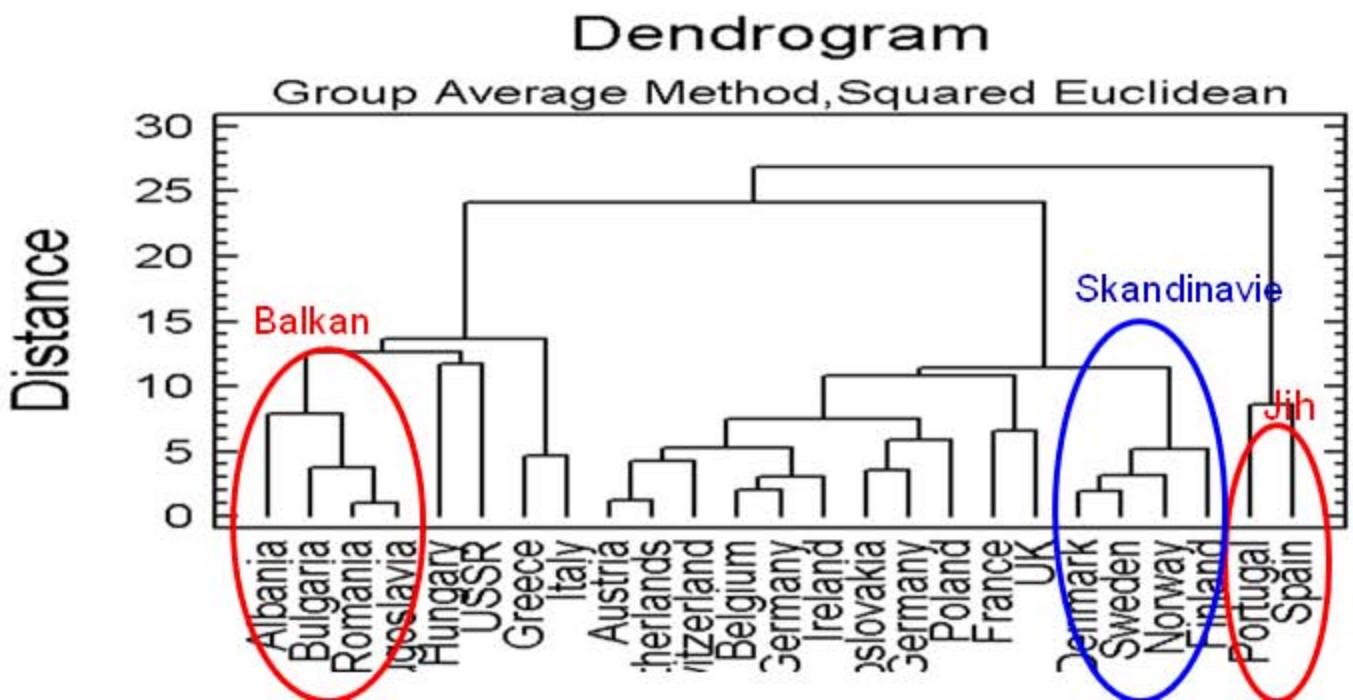
	B402x2	B402x3	B402x4
B402x3	0.991		
B402x4	0.841	0.795	
B402x5	0.845	0.852	0.836

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.



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

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

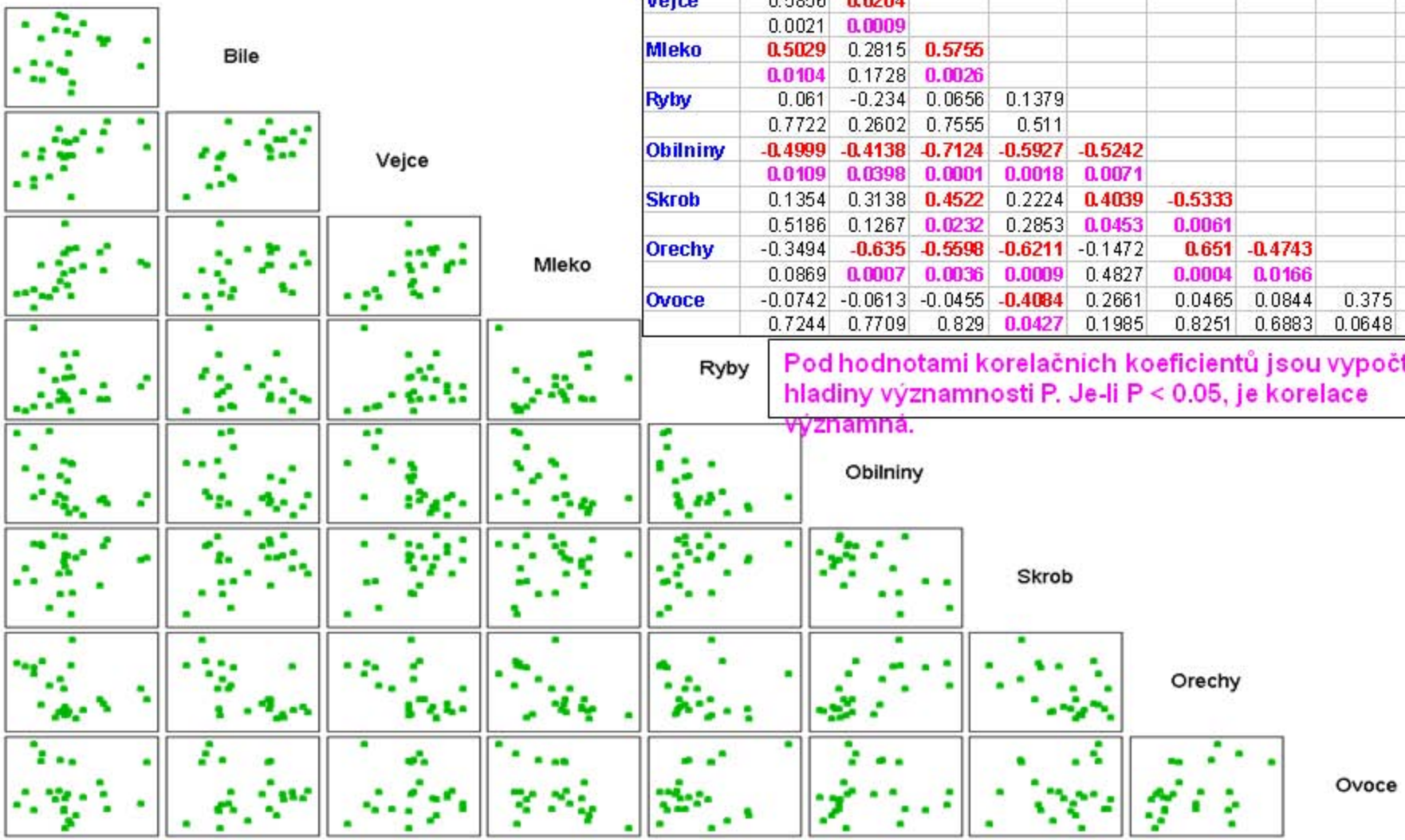
Data: *i* značí index, ***Cervene*** udává červené maso, **Bile** maso, **Vejce, Mleko**, **Ryby**, **Obilniny**, **Skrob**, **Orechy**, **Ovoce** a zelenina

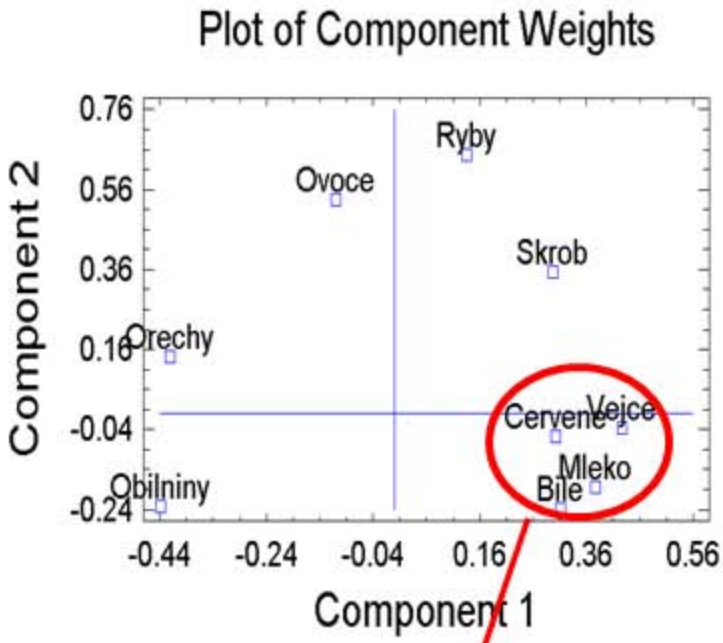
	Objekty	Proměnné								
i	**Stát**	**Cervene**	**Bile**	**Vejce**	**Mleko**	**Ryby**	**Obilniny**	**Skrob**	**Orechy**	**Ovoce**
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	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

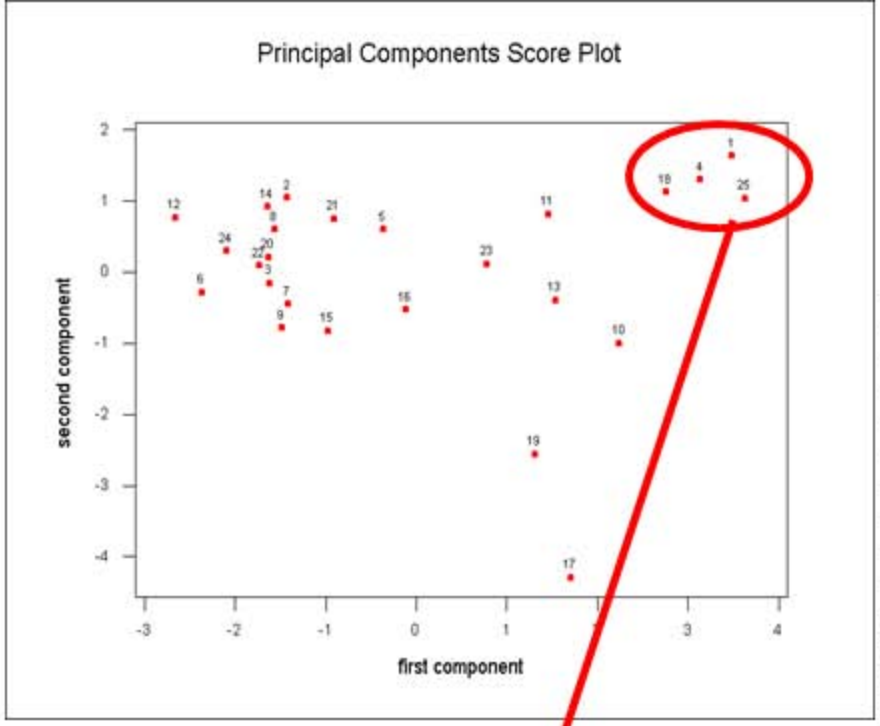
Test významnosti korelace v korelační matici

	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	

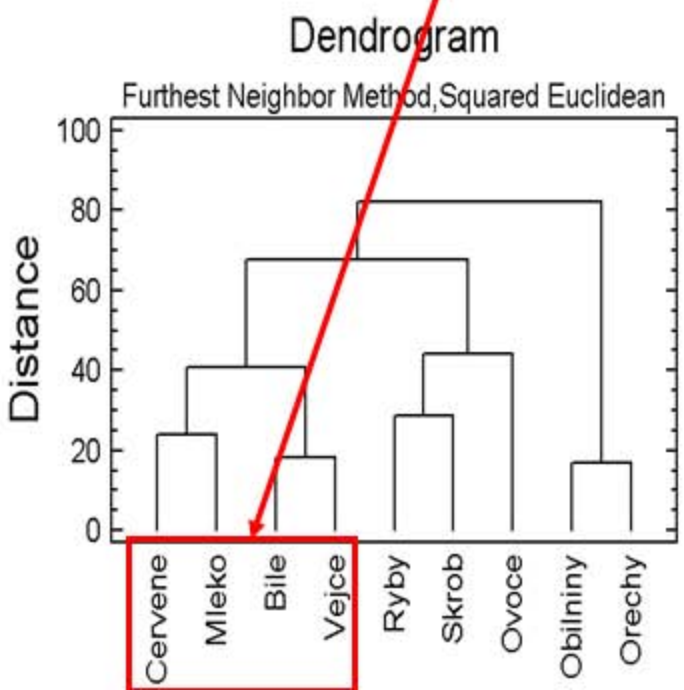
Pod hodnotami korelačních koeficientů jsou vypočtené hladiny významnosti P. Je-li P < 0.05, je korelace významná.

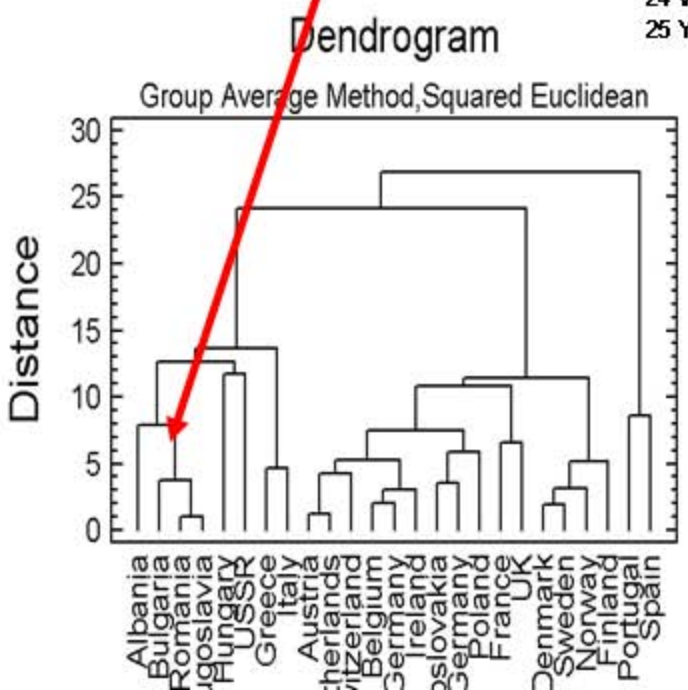


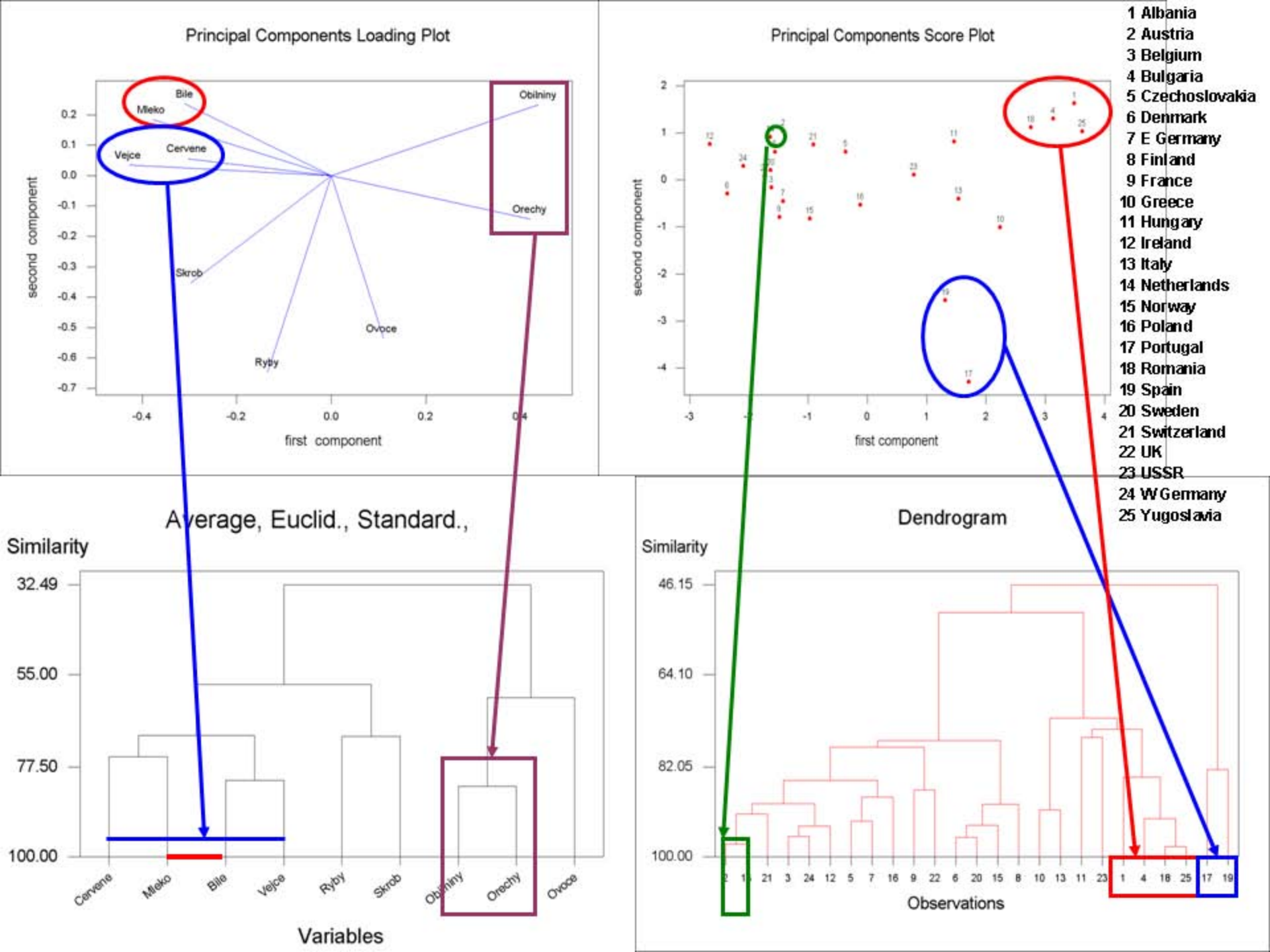




1 Albania
2 Austria
3 Belgium
4 Bulgaria
5 Czechoslovakia
6 Denmark
7 E Germany
8 Finland
9 France
10 Greece
11 Hungary
12 Ireland
13 Italy
14 Netherlands
15 Norway
16 Poland
17 Portugal
18 Romania
19 Spain
20 Sweden
21 Switzerland
22 UK
23 USSR
24 W Germany
25 Yugoslavia



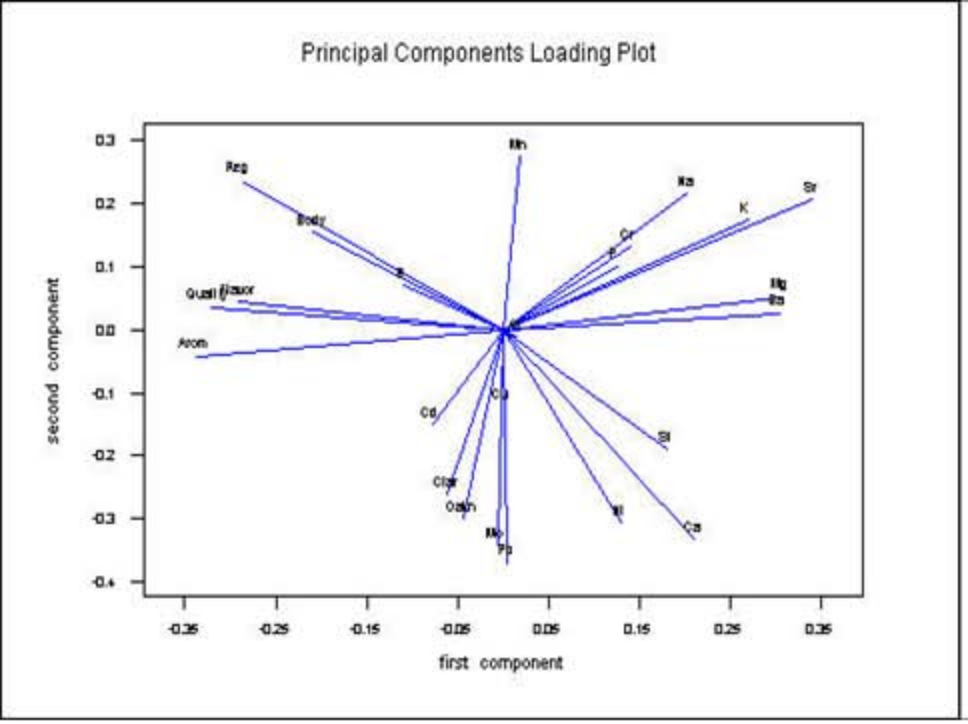


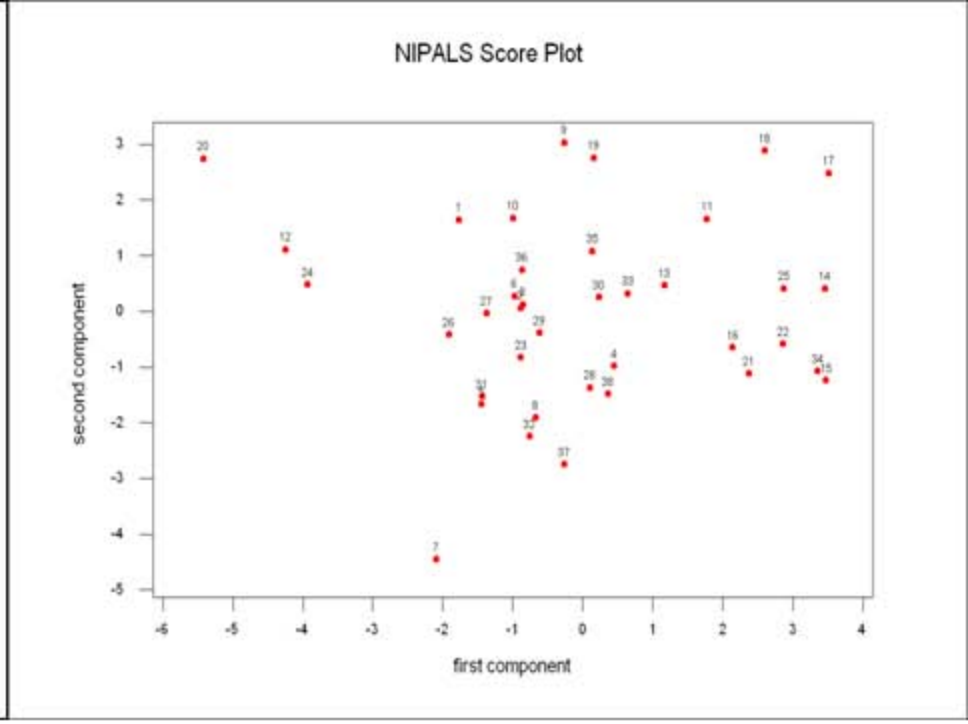

Principal Components Loading Plot
Bile
Mleko
Vejce
Cervene
Obilniny
Orechy
Skrob
Ovoce
Ryby
second component
first component
Principal Components Score Plot
second component
first component
1 Albania
2 Austria
3 Belgium
4 Bulgaria
5 Czechoslovakia
6 Denmark
7 E Germany
8 Finland
9 France
10 Greece
11 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
Average, Euclid., Standard.,
Similarity
32.49
55.00
77.50
100.00
Cervene
Mleko
Bile
Vejce
Ryby
Skrob
Obilniny
Orechy
Ovoce
Variables
Dendrogram
Similarity
46.15
64.10
82.05
100.00
Observations

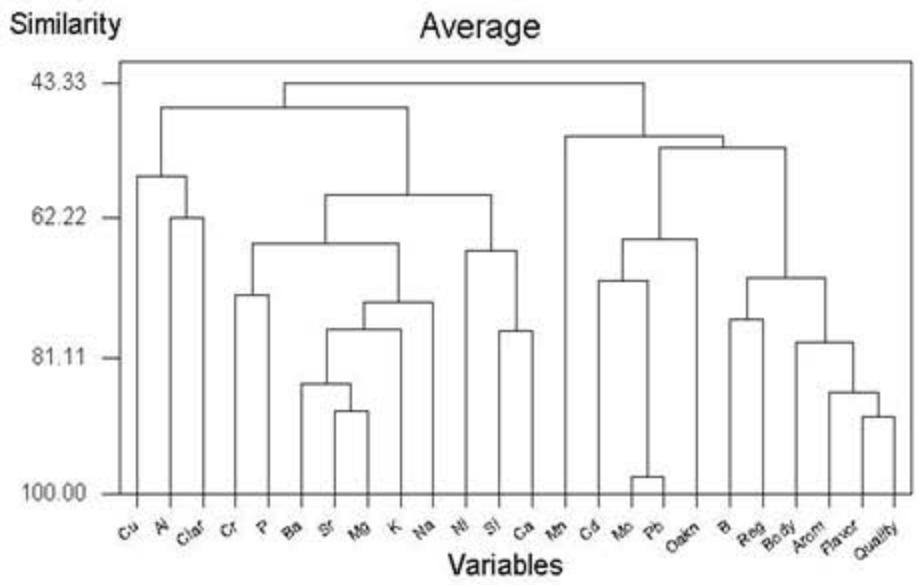
Úloha 4. Faktorová analýza při klasifikaci vzorků vín (Kompendium E408)

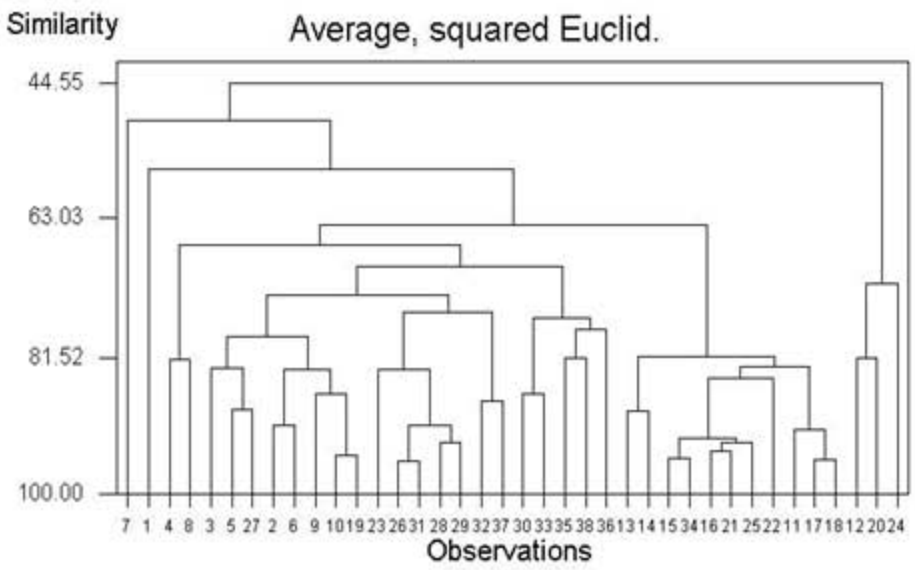
Pro 38 vzorků vín bylo nalezeno 24 analytických obsahů stopových prvků a charakteristických fyzikálně-chemických vlastností. Utvořte shluky podobných vlastností a dále shluky podobných vín.

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

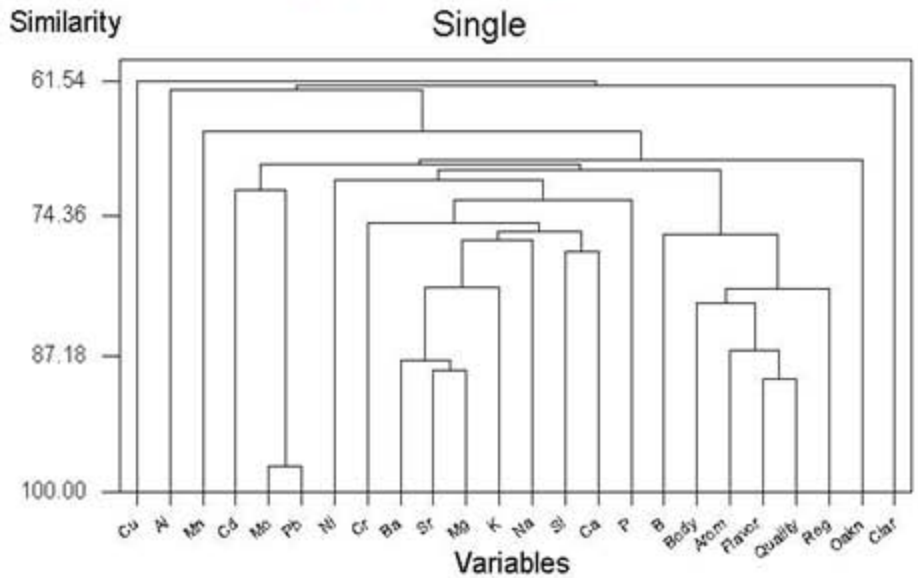

Principal Components Loading Plot
second component
first component

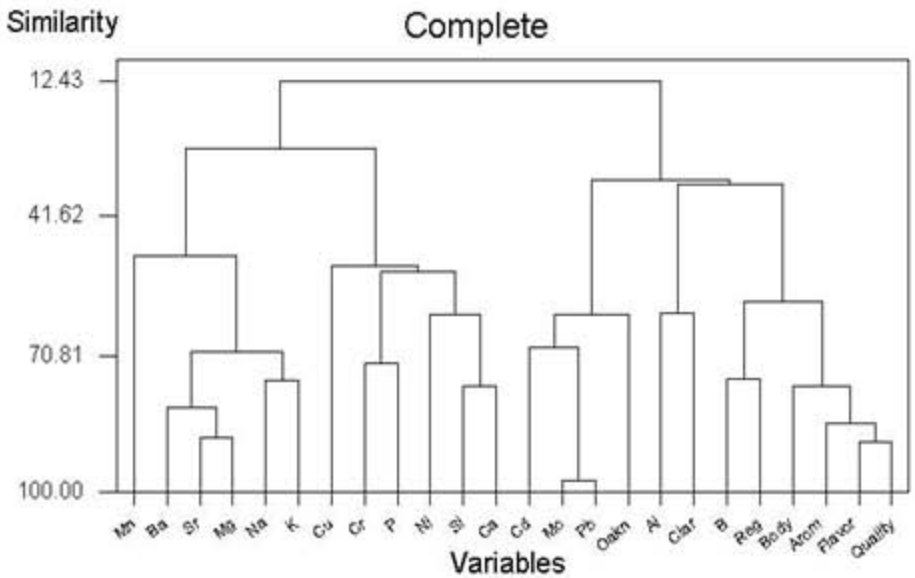

NIPALS Score Plot
second component
first component

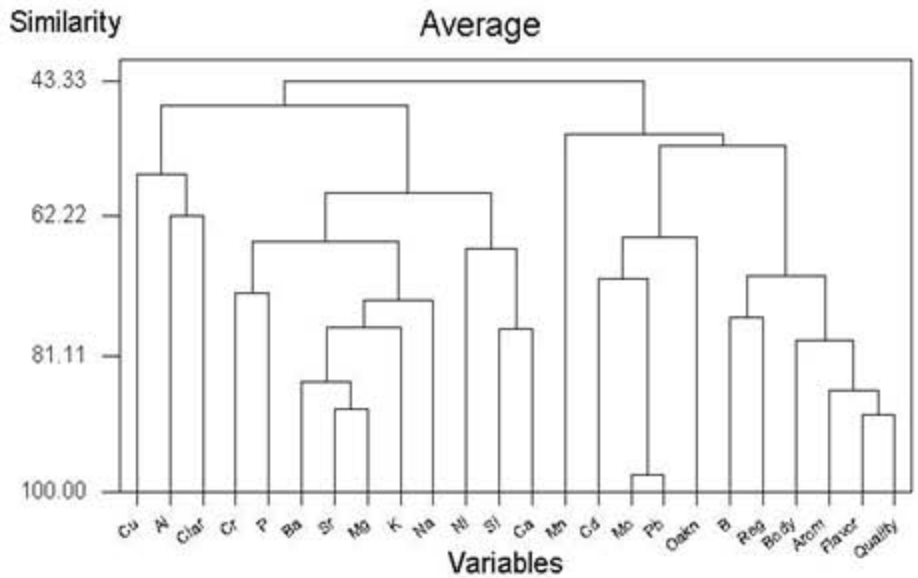

Similarity
Average
43.33
62.22
81.11
100.00
Cu Al Clar Cr P Ba Sr Mg K Na Ni Si Ca Mn Cd Mo Pb Oakn B Red Body Arom Flavor Quality
Variables

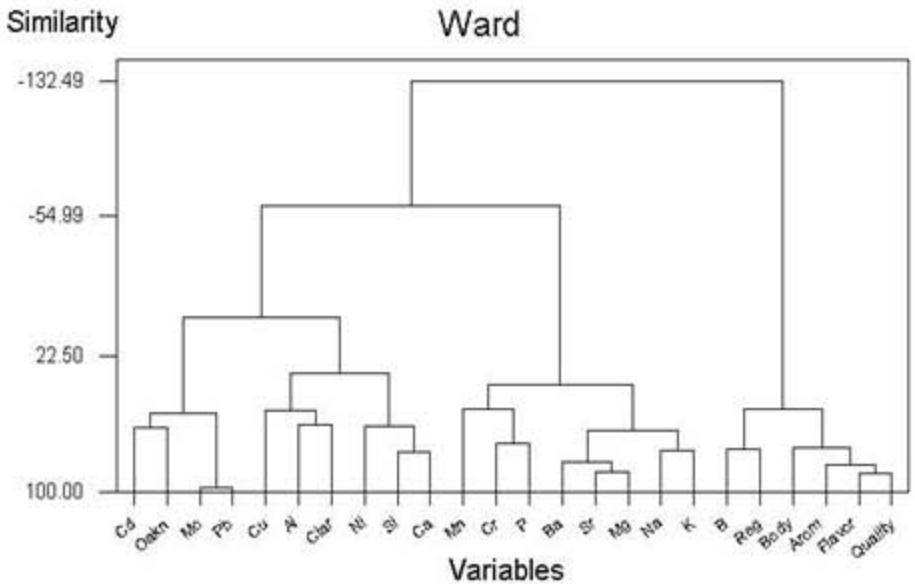

Similarity
Average, squared Euclid.
44.55
63.03
81.52
100.00
7 1 4 8 3 5 27 2 6 9 10 19 23 26 31 28 29 32 37 30 33 35 38 36 13 14 15 34 16 21 25 22 11 17 18 12 20 24
Observations

Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.

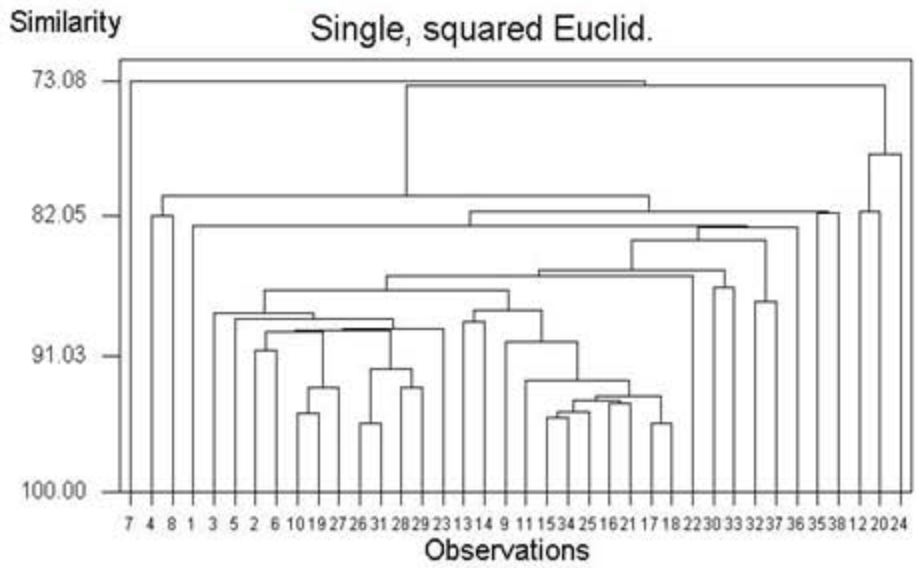


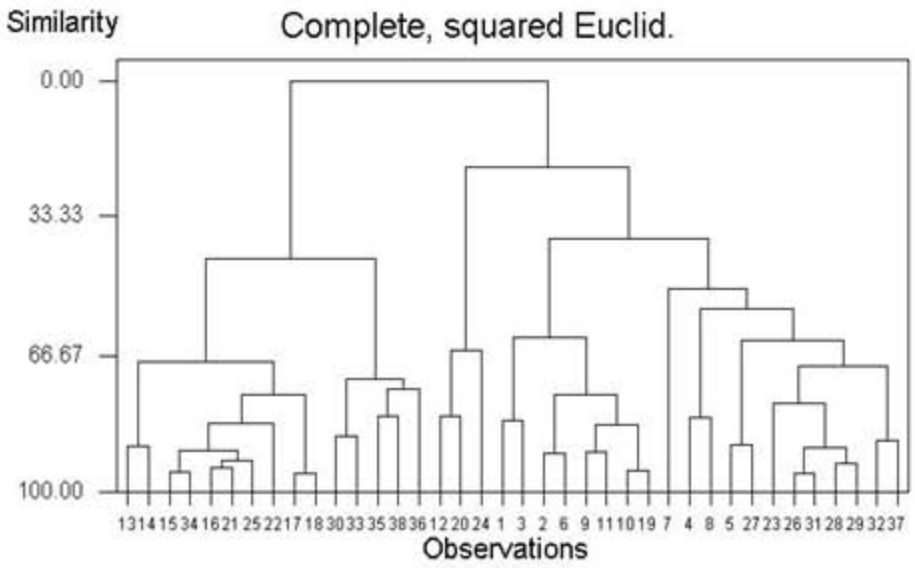


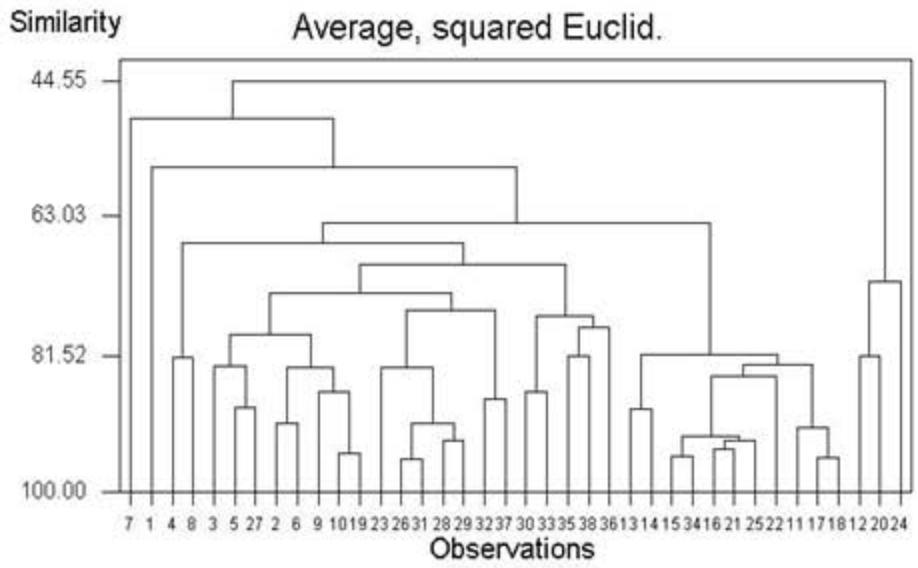


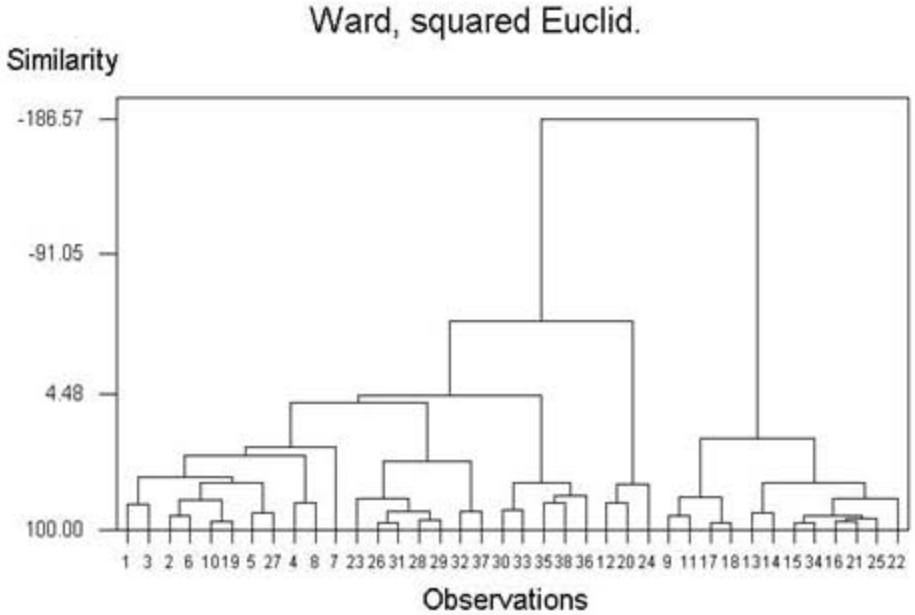


Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.







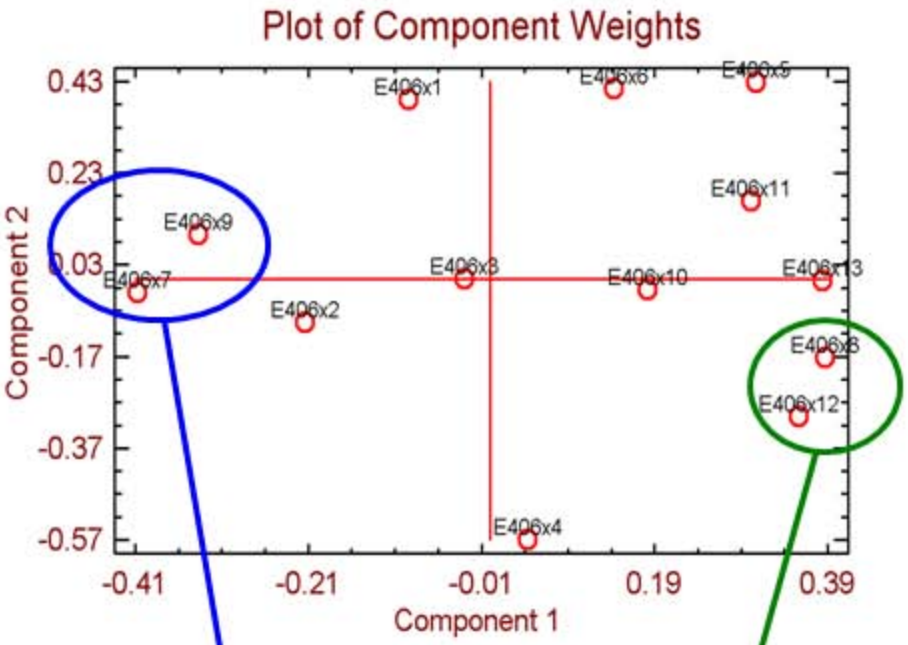


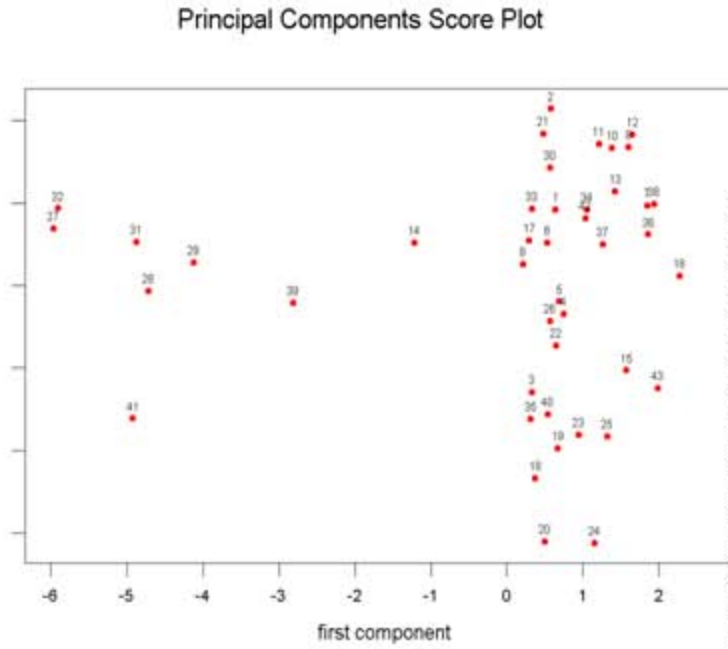
Úloha 6. Klasifikace vlastností rozličných druhů kávy (Kompendium E406)

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

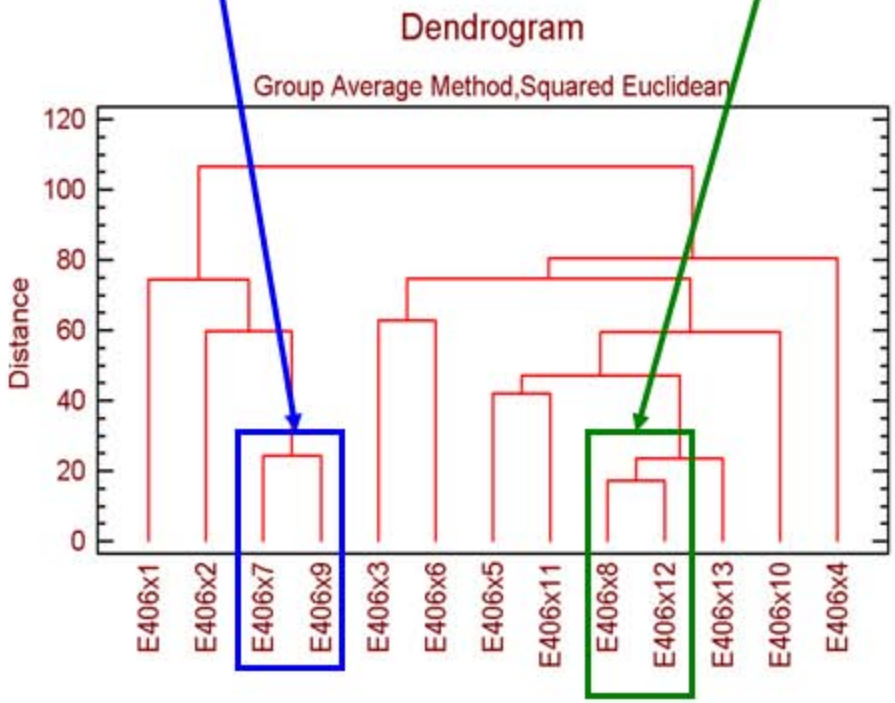
Data: 13 proměnných (sloupce): **i** index kávy, **j** je původ kávy, **x1** obsah vody, **x2** hmotnost zrn, **x3** extrakt, **x4** pH, **x5** volná acidita, **x6** obsah minerálů, **x7** tuky, **x8** kofein, **x9** trinonelin, **x10** kyselina chlorogeniková, **x11** kyselina neochlorogeniková, **x12**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

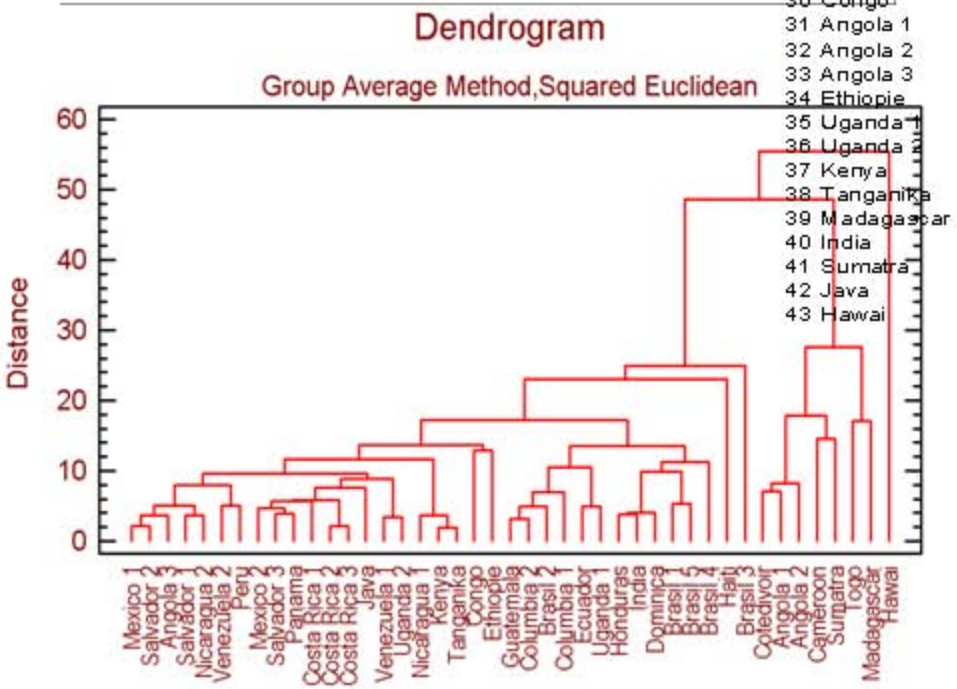
I	II	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1	5.4	0.4	0.8	6.6
2	Mexico 2	7.4	157.3	32.1	5.8	30.8	3.7	15	1.3	1	5.1	0.3	1	6.4
3	Guatemala	9.7	152.9	33.1	5.3	36.7	4.2	16.1	1.2	1	5.9	0.2	0.8	6.9
4	Honduras	10.4	174	31.5	5.6	34.2	3.9	15.8	1.1	0.9	5.9	0.4	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	Cotedlvoir	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	6.1	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	159.1	30.3	5.6	31.5	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	5.6	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	0.5	0.3	6.5





1 Mexico 1
2 Mexico 2
3 Guatemala
4 Honduras
5 Salvador 1
6 Salvador 2
7 Salvador 3
8 Nicaragua 1
9 Nicaragua 2
10 Costa Rica 1
11 Costa Rica 2
12 Costa Rica 3
13 Panama
14 Haiti
15 Dominica
16 Venezuela 1
17 Venezuela 2
18 Columbia 1
19 Columbia 2
20 Ecuador
21 Peru
22 Brasil 1
23 Brasil 2
24 Brasil 3
25 Brasil 4
26 Brasil 5
27 Cotedivoir
28 Togo
29 Cameroon
30 Congo
31 Angola 1
32 Angola 2
33 Angola 3
34 Ethiopie
35 Uganda 1
36 Uganda 2
37 Kenya
38 Tanganika
39 Madagascar
40 India
41 Sumatra
42 Java
43 Hawai





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.

Úloha 2. Vytvoření dendrogramu objektů neuroleptika (Kompendium B402)

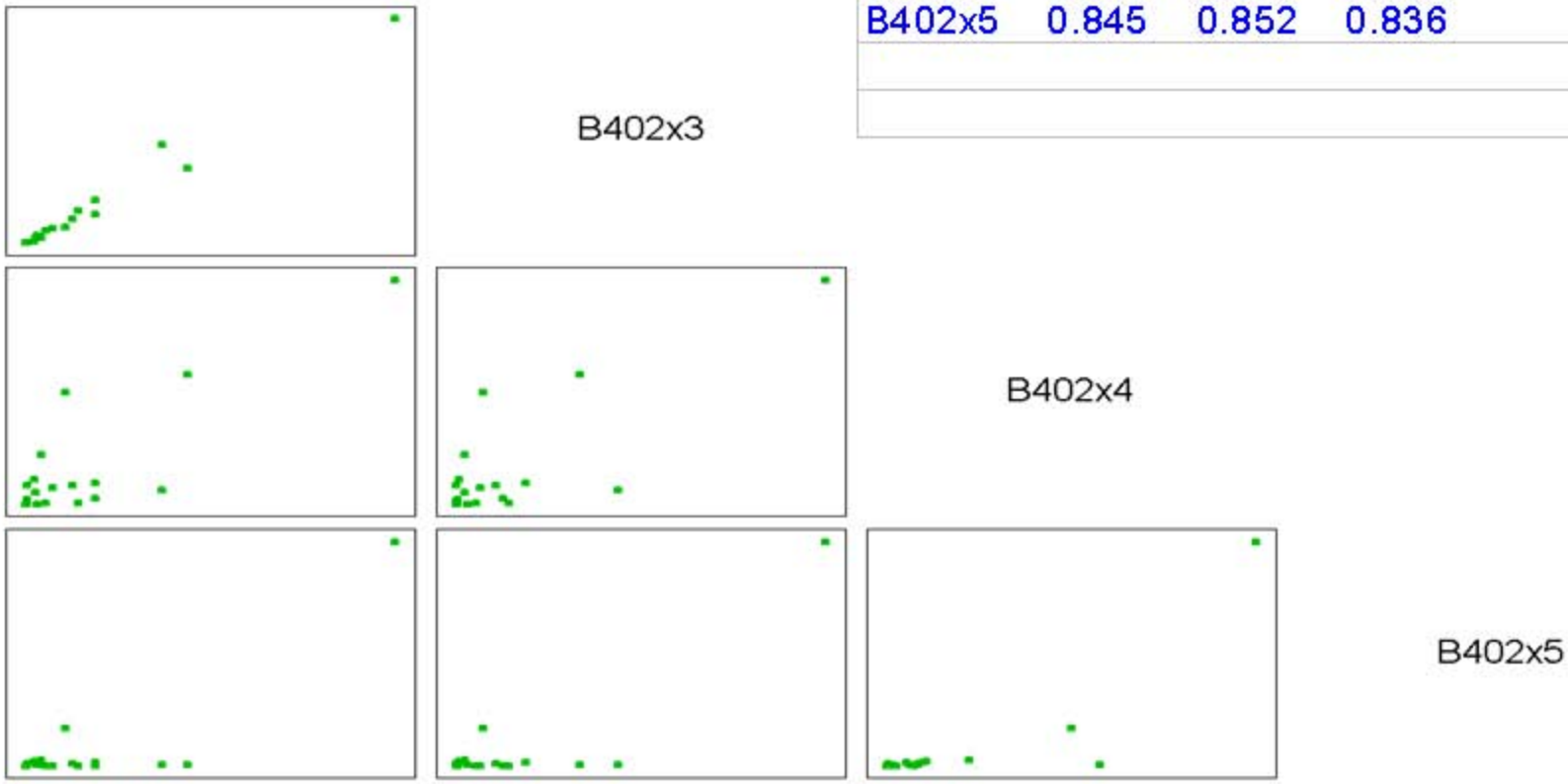
Liší se v účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění. Provedeme klasifikaci neuroleptik do shluků podobných účinků s ohledem na 4 znaky.

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

Příklad 9.4/str. 284

NCSS 2007 Data - [E:\Moje\Učebnice 2005 M+M+H\CD v knížce\DATA\NCSS2002\51Neuroleptika.S0]

1 Chlorphromazine

	Lek	Nervoz	Stereo	Tres	Usmr	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
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	0.14													
4	4 Fluphenazine																	
5	5 Perphenazine																	
6	6 Thioridazine																	
7	7 Pifluthixol																	
8	8 Thiothixene																	
9	9 Chorprothixene																	
10	10 Spiperone																	
11	11 Haloperidol																	
12	12 Azaperone																	
13	13 Pipamperone																	
14	14 Pimozide																	
15	15 Metitepine																	
16	16 Clozapine																	
17	17 Perlapine																	
18	18 Sulpiride																	
19	19 Butaclamol																	
20	20 Molindone																	

NCSS: Hierarchical Clustering / Dendrograms

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: Nervoz-Usmr

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Simple Average (Weighted Pair-Group)

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: Lek

Input Format: Raw Data

INPUT FORMAT:
Specify the format of the input dataset. Your choices are:

--RAW DATA
The variables are in the standard format in which each row represents an object and each column represents a variable.

--DISTANCES
The variables containing the distance matrix are specified in the Interval Variables option.

Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

--CORRELATIONS 1
The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:
Dij = (1-Rij)/2

--CORRELATIONS 2
The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:
Dij = 1-|Rij|

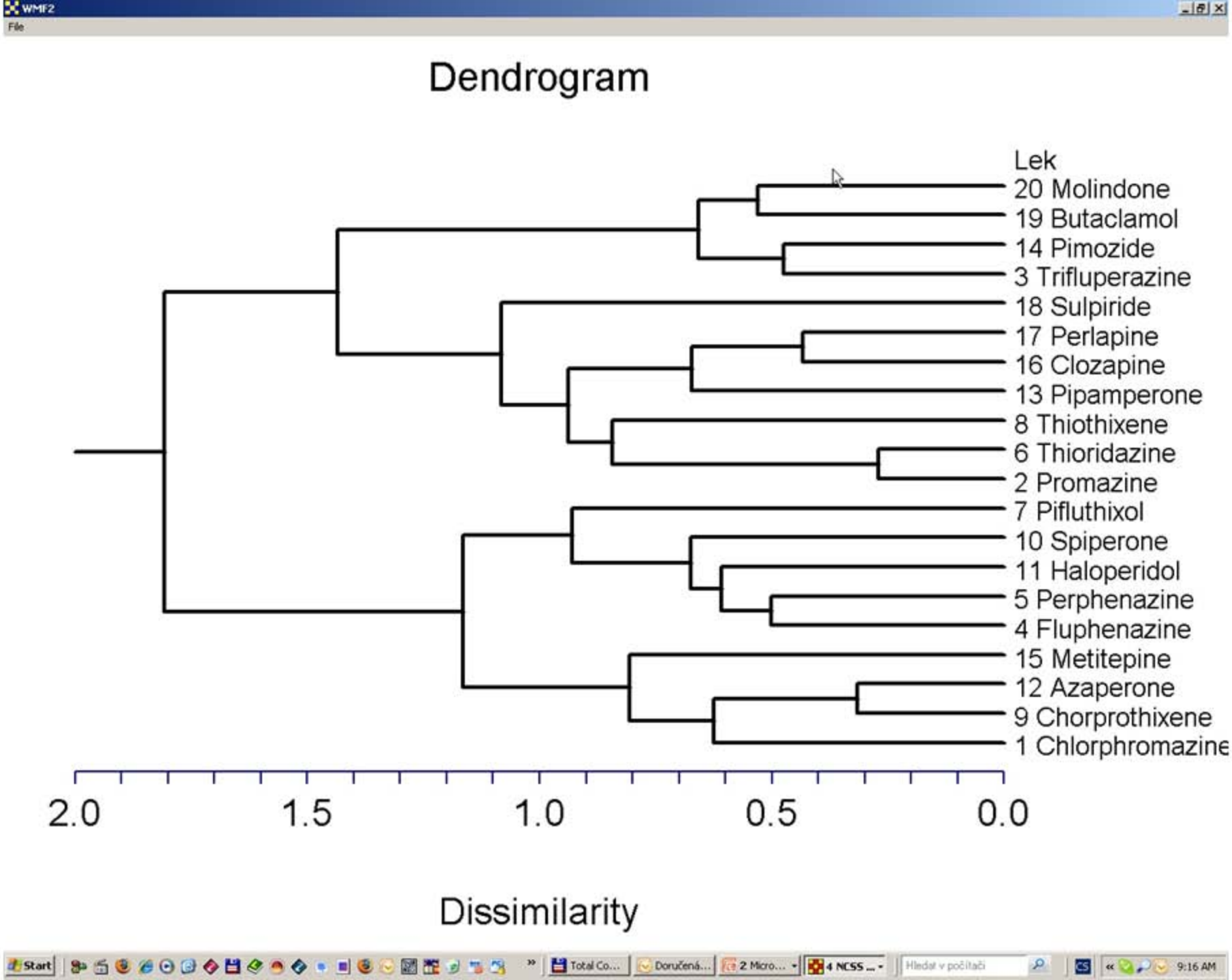
--CORRELATIONS 3
The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:
Dij = 1 - Rij*Rij

Reset | Guide Me

Opt 0 Template Id:

Variable Info | Sheet1

This is the spreadsheet that lets you enter and edit your data.


WMF2
File
Dendrogram
Lek
20 Molindone
19 Butaclamol
14 Pimozide
3 Trifluperazine
18 Sulpiride
17 Perlapine
16 Clozapine
13 Pipamperone
8 Thiothixene
6 Thioridazine
2 Promazine
7 Pifluthixol
10 Spiperone
11 Haloperidol
5 Perphenazine
4 Fluphenazine
15 Metitepine
12 Azaperone
9 Chorprothixene
1 Chlorphromazine
2.0
1.5
1.0
0.5
0.0
Dissimilarity
Start
Total Co...
Doručená...
2 Micro...
4 NCSS ...
Hledat v počítači
9:16 AM

Příklad 9.4/str. 284

NCSS Output - [NCSS: Hierarchical Clustering / Dendrograms Output]

17	18	0.822405	1.081651	-0.259246	-31.52
17	19	0.897885	1.434985	-0.537100	-59.82
17	20	0.742779	1.434985	-0.692205	-93.19
18	19	1.622020	1.434985	0.187035	11.53
18	20	1.267499	1.434985	0.167407	13.21
19					

NCSS: Hierarchical Clustering / Dendrograms

File Run Analysis Graphics Tools Window Help

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: Nervoz-Usme

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Single Linkage (Nearest Neighbor)

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: Lek

Input Format: Raw Data

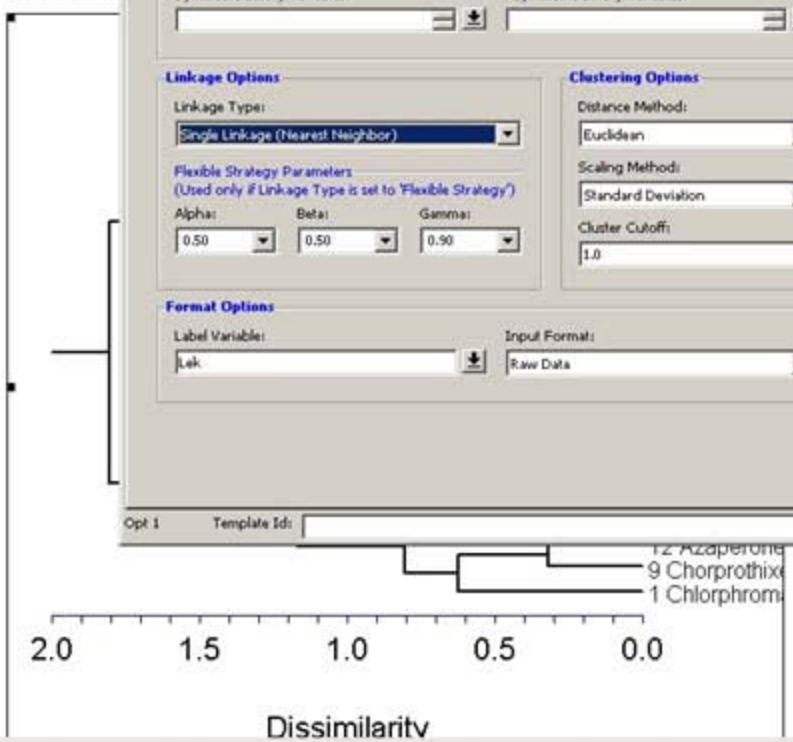
Load a printable copy of the documentation for this module.
An Acrobat PDF file will be loaded for display and printing.

Reset | Guide Me

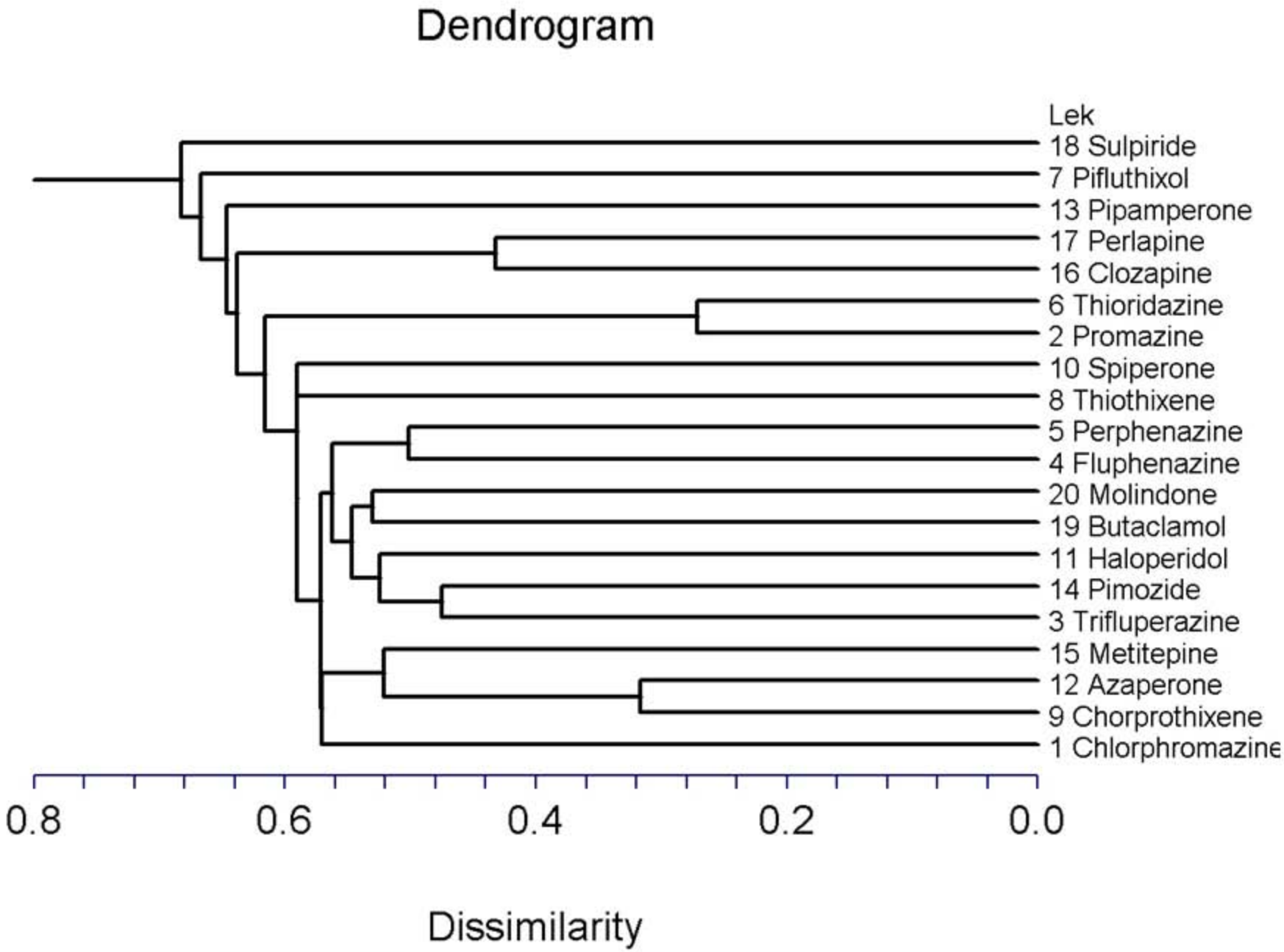
Opt 1 Template Id:

Page/Date/Tim
Database
Variables
Clustering Metl
Distance Type
Scale Type

Dendrogram



Page 7/7 Line 335 Col 1


Dendrogram
Lek
18 Sulpiride
7 Pifluthixol
13 Pipamperone
17 Perlapine
16 Clozapine
6 Thioridazine
2 Promazine
10 Spiperone
8 Thiothixene
5 Perphenazine
4 Fluphenazine
20 Molindone
19 Butaclamol
11 Haloperidol
14 Pimozide
3 Trifluperazine
15 Metitepine
12 Azaperone
9 Chorprothixene
1 Chlorphromazine
0.8
0.6
0.4
0.2
0.0
Dissimilarity

Příklad 9.4/str. 284

NCSS 2007 Data - [E:\Moje\Učebnice 2005 M+M+H\CD v knižce\DATA\NCSS2002\51Neuroleptika.S0]

1 Chlorphromazine

	Lek	Nervoz	Stereo	Tres	Usmr	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
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	0.14													
4	4 Fluphenazine																	
5	5 Perphenazine																	
6	6 Thioridazine																	
7	7 Pifluthixol																	
8	8 Thiothixene																	
9	9 Chorprothixene																	
10	10 Spiperone																	
11	11 Haloperidol																	
12	12 Azaperone																	
13	13 Pipamperone																	
14	14 Pimozide																	
15	15 Metitepine																	
16	16 Clozapine																	
17	17 Perlapine																	
18	18 Sulpiride																	
19	19 Butaclamol																	
20	20 Molindone																	

NCSS: Hierarchical Clustering / Dendrograms

File Run Analysis Graphics Tools Window Help

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: Nervoz-Usmr

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Complete Linkage (Furthest Neighbor)

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: Lek

Input Format: Raw Data

INTERVAL VARIABLES:
Designates interval-type variables (if any) or the columns of the matrix if Distance or Correlation matrix input was selected in the Input Format box.

Interval variables are continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

In general, an interval should keep the same importance throughout the scale. For example, the length of time between 1905 and 1925 is the same as the length of time between 1995 and 2015.

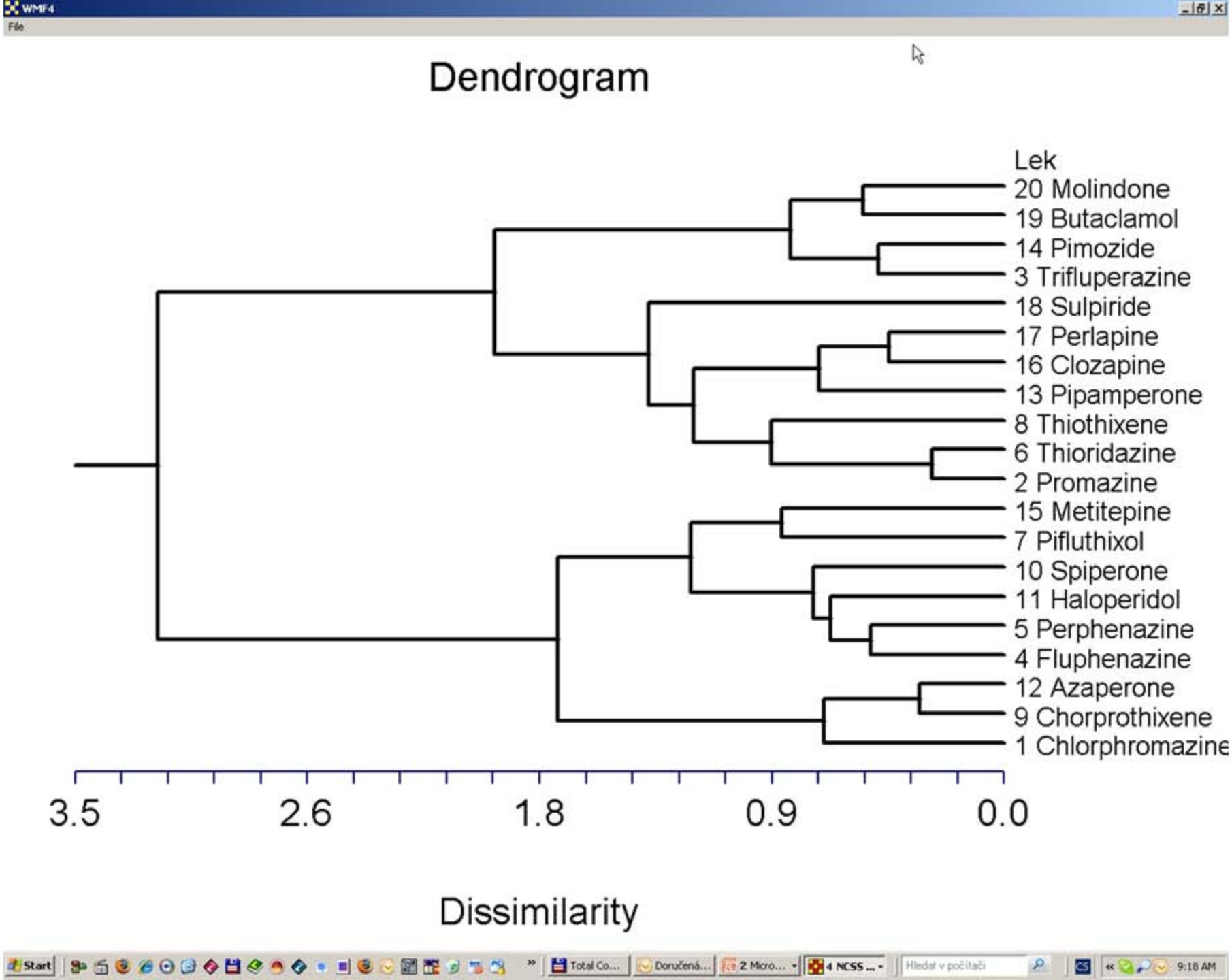
Note that a nonlinear transformation of an interval variable is probably not an interval variable. For example, the logarithm of height is not an interval variable since the value of an interval along the scale changes depending upon where you are on the scale.

Reset | Guide Me

Opt 1 Template Id:

Variable Info | Sheet1

This is the spreadsheet that lets you enter and edit your data.


WMF4
File
Dendrogram
Lek
20 Molindone
19 Butaclamol
14 Pimozide
3 Trifluperazine
18 Sulpiride
17 Perlapine
16 Clozapine
13 Pipamperone
8 Thiothixene
6 Thioridazine
2 Promazine
15 Metitepine
7 Pifluthixol
10 Spiperone
11 Haloperidol
5 Perphenazine
4 Fluphenazine
12 Azaperone
9 Chorprothixene
1 Chlorphromazine
3.5
2.6
1.8
0.9
0.0
Dissimilarity
Start
Total Co...
Doručená...
2 Micro...
4 NCSS ...
Hledat v počítači
9:18 AM

Příklad 9.4/str. 284

NCSS 2007 Data - [E:\Moje\Učebnice 2005 M+M+H\CD v knížce\DATA\NCSS2002\51Neuroleptika.S0]

1 Chlorphromazine

	Lek	Nervoz	Stereo	Tres	Usmr	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
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	0.14													
4	4 Fluphenazine																	
5	5 Perphenazine																	
6	6 Thioridazine																	
7	7 Pifluthixol																	
8	8 Thiothixene																	
9	9 Chorprothixene																	
10	10 Spiperone																	
11	11 Haloperidol																	
12	12 Azaperone																	
13	13 Pipamperone																	
14	14 Pimozide																	
15	15 Metitepine																	
16	16 Clozapine																	
17	17 Perlapine																	
18	18 Sulpiride																	
19	19 Butaclamol																	
20	20 Molindone																	

NCSS: Hierarchical Clustering / Dendrograms

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: Nervoz-Usmr

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Ward's Minimum Variance

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: Lek

Input Format: Raw Data

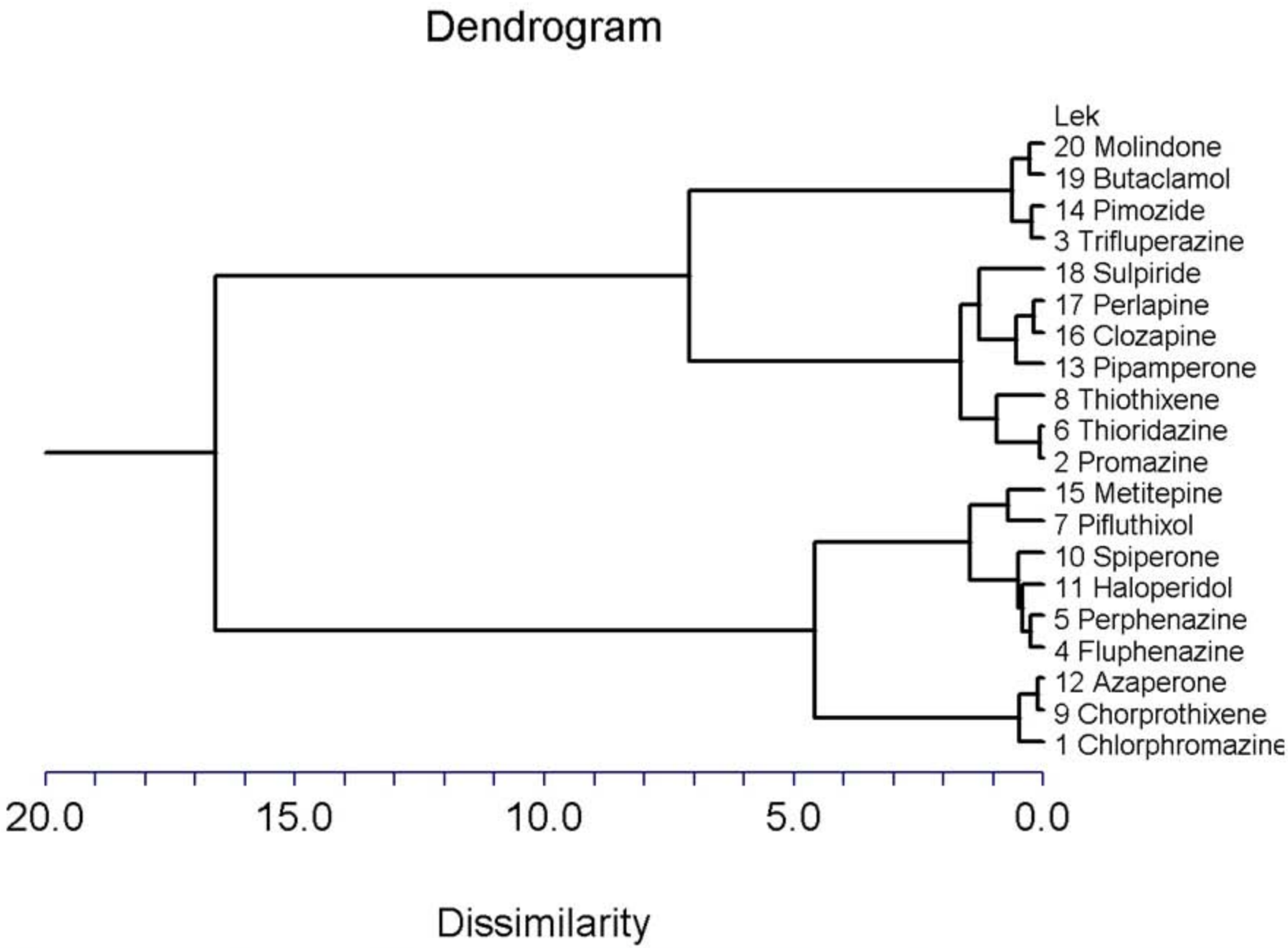
Load a printable copy of the documentation for this module.
An Acrobat PDF file will be loaded for display and printing.

Reset | Guide Me

Opt 1 Template Id:

Variable Info | Sheet1

This is the spreadsheet that lets you enter and edit your data.


Dendrogram
Lek
20 Molindone
19 Butaclamol
14 Pimozide
3 Trifluperazine
18 Sulpiride
17 Perlapine
16 Clozapine
13 Pipamperone
8 Thiothixene
6 Thioridazine
2 Promazine
15 Metitepine
7 Pifluthixol
10 Spiperone
11 Haloperidol
5 Perphenazine
4 Fluphenazine
12 Azaperone
9 Chorprothixene
1 Chlorphromazine
20.0
15.0
10.0
5.0
0.0
Dissimilarity

Příklad 9.12/str. 323

NCSS 2007 Data - [E:\Moje\Učebnice 2005 M+M+H\CD v knižce\DATA\NCSS2002\P913Kava.S0]

File Edit Data Analysis Graphics Tools Window Help

	Objekt	voda	zrno	extrakt	pH	acidita	mineral	tuky	kofein	trinonelin	kchloroge	kneochlor	kisochlor	sumakyse	C15	C16	C17	C18	C19
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													6.8					
5	Salvador 1													6.3					
6	Salvador 2													6.4					
7	Salvador 3													6.9					
8	Nicaragua													6.9					
9	Nicaragua													6.7					
10	Costa Rica													6.1					
11	Costa Rica													6.3					
12	Costa Rica													6.5					
13	Panama													6.6					
14	Haiti													7.5					
15	Dominica													6.5					
16	Venezuela													6.2					
17	Venezuela													7.3					
18	Columbia 1													6.7					
19	Columbia 2													6.9					
20	Ecuador													6.6					
21	Peru													7.3					
22	Brasil 1													6.4					
23	Brasil 2													6.2					
24	Brasil 3													6					
25	Brasil 4													6					
26	Brasil 5													6.5					
27	Cotedivoir													6.5					
28	Togo													7.1					
29	Cameroon													7.6					
30	Congo													6.4					
31	Angola 1													7.9					
32	Angola 2													8.3					
33	Angola 3													6.4					
34	Ethiopie													6.5					
35	Uganda 1													6.9					
36	Uganda 2													6.2					
37	Kenya													6					
38	Tanganika													6.9					
39	Madagasc													6.7					
40	India													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	5.6	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	0.5	0.3	6.5					
44																			
45																			
46																			

Variable Info | Sheet1

Recalculate the transformations for all variables, proceeding from left to right.

NCSS: Hierarchical Clustering / Dendrograms

File Run Analysis Graphics Tools Window Help

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: VODA-SUMAKYSEL

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Ward's Minimum Variance

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: OBJEKT

Input Format: Raw Data

VARIABLE SELECTION BUTTON:
Press this button to display a Variable selection window. You can also double-click directly in the box to the left of this button or simply type the name(s) or column number(s) in that box.

COMMENTS:
There are several ways to specify variables:

1. You can double click in the box to the left or single click this button to bring up a special Variable Selection window.

2. You can enter variable names or column position numbers directly in this box without using the Variable selection window. When entering variables this way, you can enter a range of columns such as 1, 2, 4-7 or X1, X3-X9, X14-X20.

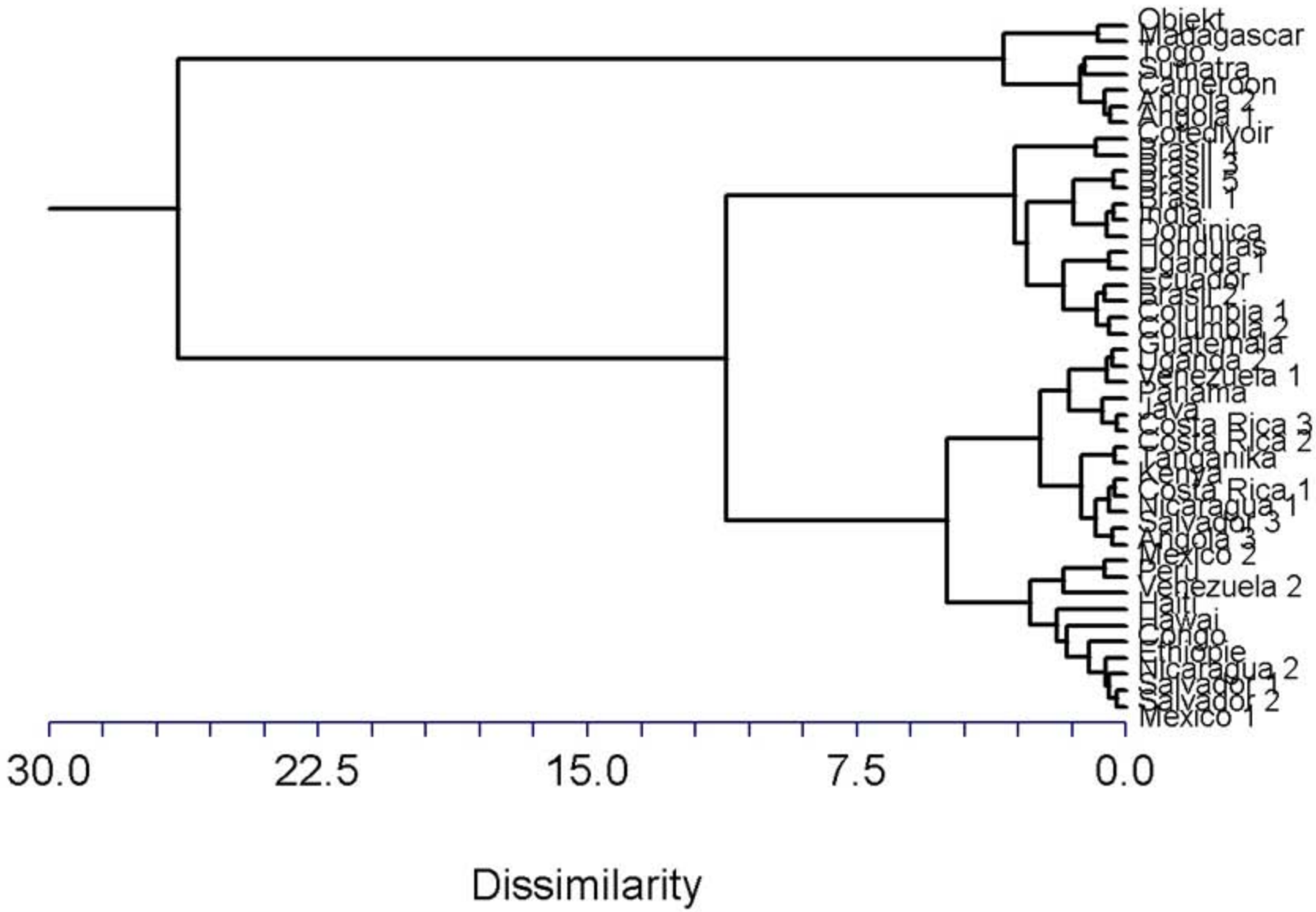
Reset | Guide Me

Opt 7 Template Id:

Start | Total Co... | Doručená... | 2 Micro... | 3 NCSS ... | Hledat v počítači | 11:17 AM

WMF8
File

Dendrogram



Dissimilarity

Start
Total Co...
Doručená...
2 Micro...
4 NCSS ...
Hledat v počítači
11:17 AM

Příklad 9.13/str. 325

NCSS 2007 Data - [E:\Moje\Učebnice 2005 M+M+H\CD v knižce\DATA\NCSS2002\P914Vina.S0]

File Edit Data Analysis Graphics Tools Window Help

Olo0171

	Objekt	kateg	alkohol	necuk	fosfaty	fenoly	flavan	pomerA1	pomerA2	prolin	C11	C12	C13	C14	C15	C16	C17	C18	C19
1	Olo0171	1	14.23	24.82	320	2.8	3.06	3.92	4.77	1065									
2	Olo0.371	1	13.16	26.3	497	2.8	3.24	3.17	3.46	1185									
3	Olo0571	1	13.24	26.05	408	2.8	2.69	2.93	3.22	735									
4	Olo0771																		
5	Olo0971																		
6	Olo1171																		
7	Olo1371																		
8	Olo1571																		
9	Olo1771																		
10	Olo1971																		
11	Olo0273																		
12	Olo0473																		
13	Olo0673																		
14	Olo0873																		
15	Olo1073																		
16	Olo1273																		
17	Olo1473																		
18	Olo1673																		
19	Olo1873																		
20	Olo2073																		
21	Olo0274																		
22	Olo0474																		
23	Olo0674																		
24	Olo0874																		
25	Olo1074																		
26	Olo1274																		
27	Olo1474																		
28	Olo1674																		
29	Olo1874																		
30	Olo2074																		
31	Gri0270																		
32	Gri0470																		
33	Gri0670																		
34	Gri0970																		
35	Gri0271																		
36	Gri0471																		
37	Gri0671																		
38	Gri0971																		
39	Gri0272																		
40	Gri0472																		
41	Gri0672	2	12.08	23.5	215	2	1.58	2.72	3.5	630									
42	Gri0173	2	11.84	23.4	378	2.2	2.21	3.08	3.81	520									
43	Gri0373	2	12.16	25.8	285	1.78	1.69	2.26	2.92	495									
44	Gri0673	2	12.08	24	278	2.2	1.59	3.21	3.77	625									
45	Gri0873	2	12	25.3	342	1.45	1.25	2.65	2.75	450									
46	Gri0174	2	12.29	25	329	2.45	2.25	3.3	3.75	290									

Variable Info | Sheet1

This is the spreadsheet that lets you enter and edit your data.

NCSS: Hierarchical Clustering / Dendrograms

File Run Analysis Graphics Tools Window Help

Variables | Reports | Dendrogram | Storage | Template

Variables

Interval Variables:

Ratio Variables:

Ordinal Variables: alkohol-prolin

Nominal Variables:

Symmetric Binary Variables:

Asymmetric Binary Variables:

Linkage Options

Linkage Type: Ward's Minimum Variance

Flexible Strategy Parameters
(Used only if Linkage Type is set to 'Flexible Strategy')

Alpha: 0.50 Beta: 0.50 Gamma: 0.90

Clustering Options

Distance Method: Euclidean

Scaling Method: Standard Deviation

Cluster Cutoff: 1.0

Format Options

Label Variable: Objekt

Input Format: Raw Data

QUICK LAUNCH:
Press to see a list of available procedures.

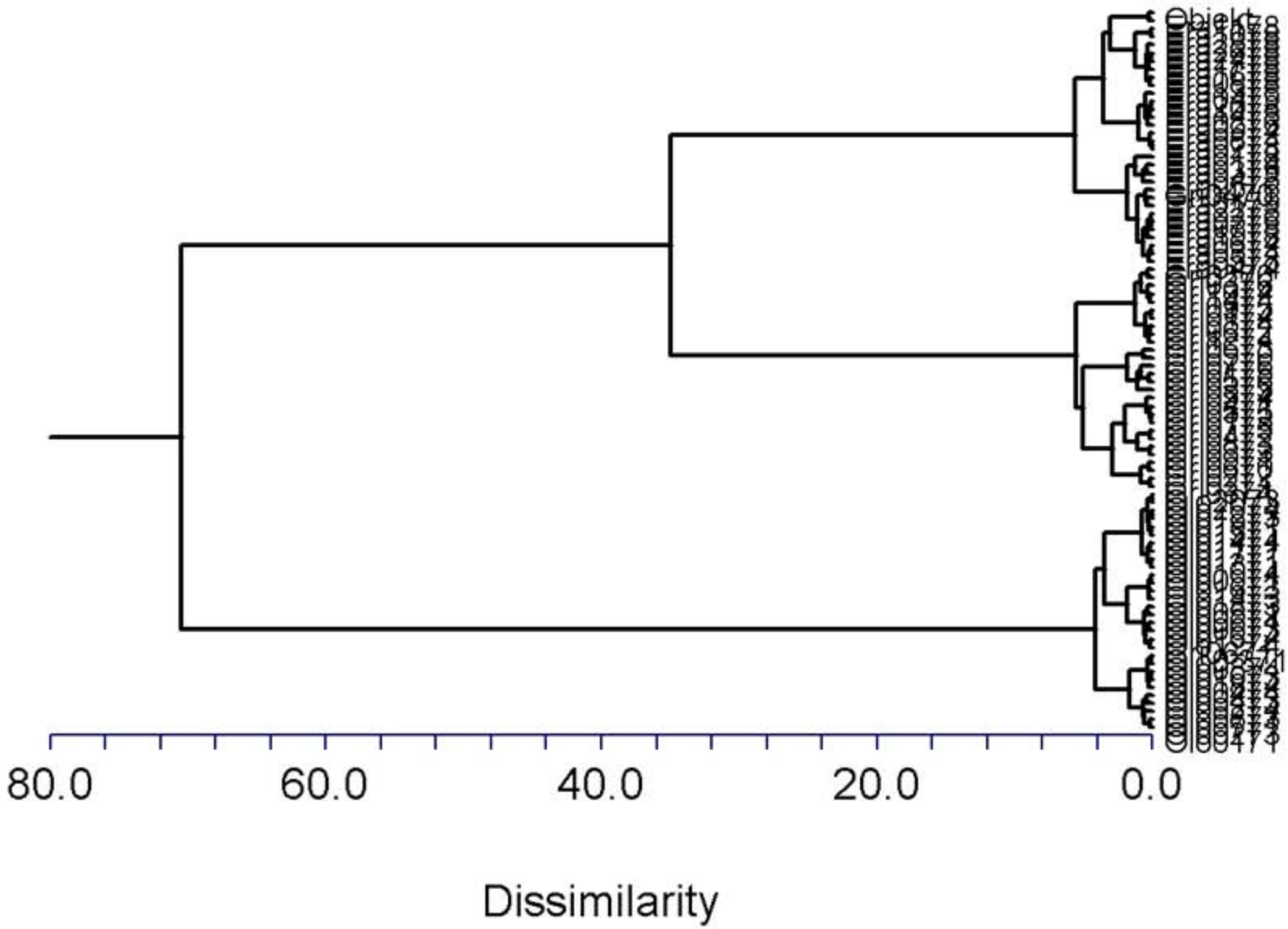
Reset | Guide Me

Opt 6 Template Id:

Start | Total Co... | Doručená... | 2 Micro... | 2 NCSS ... | Hledat v počítači | 11:19 AM

WMF9
File

Dendrogram



Start
Total Co...
Doručená...
2 Micro...
4 NCSS ...
Hledat v počítači
11:20 AM