Přednosti analýzy shluků ve vícerozměrné statistické analýze

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice, milan.meloun@upce.cz
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec, jiri.militky@vslib.cz

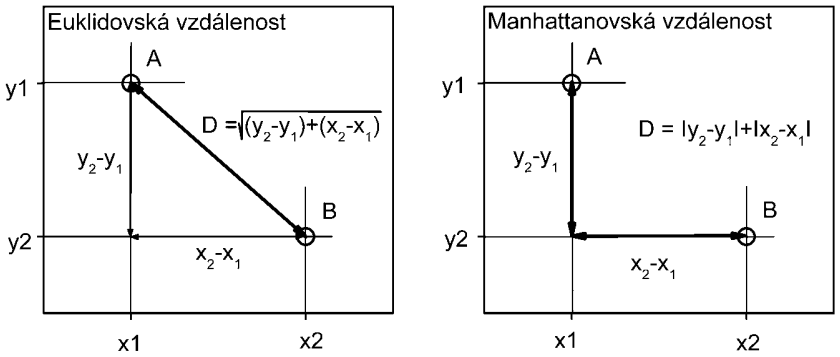
Souhrn: *Analýza shluků (Cluster analysis) patří mezi metody, které se zabývají vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství znaků) a jejich klasifikací do tříd čili shluků. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. V biologii se proto užívá shluková analýza ke klasifikování živočichů a rostlin. Tato klasifikace se nazývá numerická taxonomie. V medicině analýza shluků identifikuje nemoci a jejich stádia. Lze formulovat tři hlavní cíle analýzy shluků: popis systematiky, je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů, zjednodušení dat, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty, a konečně identifikace vztahu, kdy po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty.*

Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková analýza neobsahuje techniku k rozlišení významných a nevýznamných znaků. Provede pouze odlišení shluků. Nesprávné zařazení znaků vede k zahrnutí i odlehlých objektů, které mohou mít rušivý vliv na výsledky analýzy. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

Identifikace odlehlých objektů: Při odhalování struktury objektů je shluková analýza velmi citlivá na přítomnost nevýznamných znaků. Je citlivá také na přítomnost odlehlých objektů, které se silně odlišují ode všech ostatních objektů. Odlehlé objekty mohou předsta-vovat buď (1) skutečně odchýlené, patologické objekty, které nejsou představiteli analyzované populace, nebo (2) chybný výběr objektu z populace, který způsobí nevhodné zastoupení původní populace.

Míry podobnosti: Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů. Nejdříve se stanovují znaky, určující podobnost, které se dále kombinují do podob-nostních měr. Tímto způsobem pak může být objekt porovnán s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin:

(1) *Korelační míry:* Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále může být *Pearsonův párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedničce. V případě ordinální škály (pořadová čísla) je analogickou mírou podobností *Spearmanův korelační koeficient.* Obyčejně se vychází z transponované matice dat $\boldsymbol{X}^T$, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice $\boldsymbol{X}^T$ představují korelaci mezi dvojicí objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou "podobnost" a nízká korelace pak "nepodobnost" profilů.



Obr. 1 Nejpoužívanější míry vzdálenosti: (a) Eukleidovská *D*, (b) Manhattanovská vzdálenost *D*

(2) *Míry vzdálenosti:* Představují nejčastěji užívané míry, založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Nejčastější vzdálenostní mírou je *Eukleidovská vzdálenost* zvaná také *geometrická metrika*, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že vzdálenost

$$d_E(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2} \qquad (1)$$

představuje standardní typ vzdálenosti. Vedle Eukleidovské vzdálenosti se užívá také *čtverec Eukleidovské vzdálenosti,* který tvoří základ Wardovy metody shlukování. Často je užívaná *Manhattanská vzdálenost* zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika*, definovaná vztahem

$$d_H(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}| \qquad (2)$$

Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. Další mírou je zobecněná *Minkovského metrika,* pro kterou platí

$$d_M(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z} \qquad (3)$$

kde pro $z = 1$ jde o Hammingovu metriku a pro $z = 2$ o Eukleidovu. Čím je z větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. V některých případech se používá také *tětivová vzdálenost* (anglicky chord distance), definovaná vztahem

$$d_{CH}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{2\left[1 - \frac{\sum_{j=1}^{m} x_{kj}\, x_{lj}}{\sum_{j=1}^{m} x^2_{kj} \sum_{j=1}^{m} x^2_{lj}}\right]} \qquad (4)$$

V případě třech znaků je tětivová vzdálenost přímou vzdáleností dvou bodů na povrchu koule s jednotkovým poloměrem a počátkem v těžišti.

Problém všech vzdálenostních měr vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, díky často veliké odlišnosti jednotek měření. Shluky různých vzdálenostních měr se budou lišit, největší rozptýlení mezi shluky bude u čtverce Eukleidovské vzdálenosti. Pořadí podobností se významně změní se změnou měřítka nebo změnou jednotek jednoho ze znaků.

Všechny dosud uvedené metriky neuvažují závislost mezi znaky. Zahrneme-li do vztahu pro vzdálenost také vazby mezi znaky, vyjádřené kovarianční maticí $\boldsymbol{C}$, dostaneme novou statistickou míru, zvanou *Mahalanobisova metrika*

$$d_{Ma}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{(\boldsymbol{x}_k - \boldsymbol{x}_l)^{\mathrm{T}}\, \boldsymbol{C}^{-1}\, (\boldsymbol{x}_k - \boldsymbol{x}_l)} \qquad (5)$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální. Vysoce korelo-vaný výběr znaků může skrytě převážit celý soubor znaků shlukování.

(3) *Míry asociace:* Míry asociace podobnosti se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru (například binární proměnné). Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*. Míra asociace pak vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti na danou otázku odpověděli *ano* nebo *ne*, tedy 1 nebo 0. Rozšíření tohoto jednoduchého "souhlasného koeficientu" je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků.

Uvedeme příklad: předpokládejme, že sledujeme asociaci mezi dvěma objekty O_i a O_j. Možné binární odezvy typu 0-1 je pak možno zapsat do tzv. kontingenční tabulky:

		Objekt O_i	
Objekt O_j		1	0
	1	a	b
	0	c	d

V této tabulce jsou shrnuty všechny možné kombinace počtu znaků pro dva objekty:*a* značí počet znaků, kde mají oba objekty O_j a O_i hodnotu 1 a jde o tzv. *pozitivní shodu,b* značí počet znaků, kde má objekt O_j hodnotu 1 a objekt O_i hodnotu 0. *c* značí počet znaků, kde má objekt O_j hodnotu 0 a objekt O_i hodnotu 1. *d* značí počet znaků, kde mají oba objekty O_j a O_i hodnotu 0 a jde o tzv. *negativní shodu.* Míry asociace pak vyjadřují relativní podíly počtu znaků s ohledem na to, zda má smysl uvažovat negativní shodu nebo zda má nulová hodnota znaku u porovnávaných objektů stejnou příčinu. Mezi základní koeficienty podobnosti potom patří:

(a) *Sokalův-Michenerův koeficient asociace* (zvaný také *koeficient jednoduché shody*)

$$S_{SM} = \frac{a + d}{a + b + c + d} \tag{6}$$

(b) *Russelův-Raoův koeficient asociace*

$$S_{RR} = \frac{d}{a + b + c + d} \tag{7}$$

(c) *Jaccardův koeficient* se liší od předešlého Russelova-Raova koeficientu asociace jen tím, že má v čitateli počet shodných znaků *a*.

(d) *Hamannův koeficient asociace*

$$S_H = \frac{a + d - b - c}{a + b + c + d} \tag{8}$$

a také lze konstruovat následující obdobu

(e) *Korelační koeficient*

$$r_B = \frac{a\,d - b\,c}{\sqrt{(a + b)\,(c + d)\,(a + c)\,(b + d)}} \tag{9}$$

(f) *Rogersův a Tanimotův koeficient asociace* je definován vztahem

$$S_{RT} = \frac{a + d}{a + 2b + 2c + d} \tag{10}$$

(g) *Sörensenův koeficient asociace* je definován vztahem

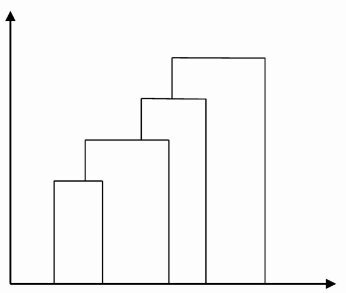
$$S_S = \frac{2a}{2a + b + c} \tag{11}$$

Standardizace dat: Před vlastní shlukovou analýzou je třeba řešit otázku, zda je třeba data standardizovat. Musí se respektovat fakt, že většina měr vzdálenosti je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají větší vliv na míru podobnosti.

(1) *Standardizování znaků*: Nejužívanější formou standardizace je *normalizace každého znaku* do svého *Z-skóre*, tj. odečtením průměru a dělením směrodatnou

odchylkou. Tato standardizace je známa pod názvem *normovací Z-funkce*. Tato transformace eliminuje rozdíly v měřítku, mnohdy i řádově se lišících znaků. Výhody standardizace znaků jsou následující: (a) Znaky lze v jednotném měřítku, (kde je průměrná hodnota 0 a směrodatná odchylka 1) vzájemně porovnávat snadněji. Kladné hodnoty jsou nad průměrem a záporné hodnoty jsou pod průměrem, (b) Se změnou měřítka nedojde k rozdílu mezi standardizovanými znaky. I když změníme například u znaku čas jednotky z minut na sekundy, standardizované hodnoty budou stále stejné.

(2) *Standardizace objektů:* Když chceme identifikovat shluky dle vzdálenosti pak standardizace není vhodná. Standardizace objektů nebo-li řádková standardizace může být však efektní ve speciálních případech.

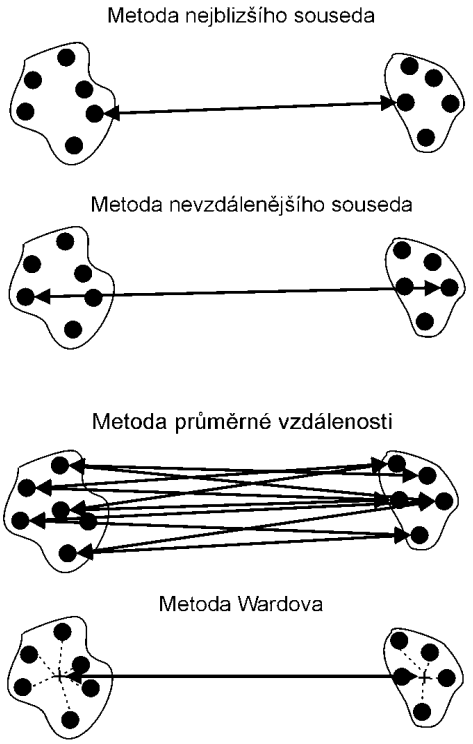


Obr. 2 Dendrogram pro metodu nejbližšího souseda

Způsoby shlukování: *Shluk* (cluster) je skupina objektů, jejichž vzdálenost je menší než vzdálenost s objekty do shluku nepatřícími. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerační* a *divizní.*

1. Hierarchické shlukovací postupy jsou založeny na hierarchickém uspořádání objektů a jejich shluků. Graficky se hierarchicky uspořádané shluky zobrazují formou *vývojového stromu* nebo *dendrogramu*. U *aglomeračního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako objekt. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Divizní postup* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Mezi metody metriky shlukování patří:

(1) *Metoda nejbližšího souseda:* Postup je postaven na minimální vzdálenosti. Naleznou se dva objekty, oddělené nejkratší vzdáleností a umístí se do shluku. Další shluk je vytvořen přidáním třetího nejbližšího objektu. Proces se opakuje až jsou všechny objekty v jednom společném shluku. Vzdálenost mezi dvěma shluky je definována jako nejkratší vzdálenost libovolného bodu v prvním shluku vůči libovolnému bodu v druhém. Dva shluky jsou propojeny v libovolném stádiu nejkratší spojkou. Typický dendrogram má pro tento případ tvar:



Obr. 3 Nejčastěji užívané metriky shlukování.

(2) *Metoda nejvzdálenějšího souseda:* Kritérium je postaveno nikoliv na

minimální ale na maximální vzdálenosti. Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích. Metoda se také nazývá *metodou úplného propojení,* protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti čili minimální podobnosti.

(3) *Metoda průměrné vzdálenosti:* Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

(4) *Wardova metoda:* Principem není optimalizace vzdáleností mezi shluky ale minimalizace heterogenity shluků podle kritéria minima přírustku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku.

(5) *Metoda těžiště:* Jde o vzdálenost dvou těžišť shluků, vyjádřených Eukleidovskou vzdáleností nebo čtvercem Eukleidovské vzdálenosti. Těžiště shluku má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé znaky. Po každém kroku shlukování se počítá nové těžiště. Poloha těžiště shluku poněkud migruje tak jak se připojují nové objekty a vznikají větší shluky.

(6) *Metoda mediánová*: Jde o jisté vylepšení metody těžiště, neboť se snaží odstranit rozdílné významnosti, které metoda těžiště dává různě velkým shlukům.

2. Nehierarchické shlukovací postupy: U *metody zárodečných bodů* (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají tvořit zárodky nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich Eukleidovské vzdálenosti od těchto typických objektů. Existuje několik postupů zadávání zárodků shluku a zařazování objektů do shluku. Těmto metodám se říká *K-means shlukování:*

(a) *Sekvenční práh:* Metoda začíná volbou jednoho zárodku shluku a zahrnuje všechny objekty uvnitř předspecifikované vzdálenosti. Když jsou všechny objekty uvnitř této vzdálenosti zahrnuty do shluku, je vybrán zárodek druhého shluku a všechny objekty uvnitř předspecifikované vzdálenosti jsou zahrnuty do shluku. Pak je vybrán třetí zárodek shluku a proces se opakuje. Když je jednou objekt shlukován se zárodkem, není s ním více počítáno do některého jiného shluku.

(b) *Paralelní práh:* Na rozdíl od předešlého postupu tento postup vybírá na začátku několik shlukových zárodků současně čili paralelně a zařazuje objekty uvnitř prahové vzdálenosti do nejbližšího zárodku. Jak se proces vyvíjí, prahovou vzdálenost lze nastavit tak, aby zařadila více nebo méně objektů do shluku. Některé objekty mohou zůstat nezařazeny do shluků, když se totiž nacházejí vně předspecifikované vzdálenosti od shlukového zárodku.

(c) *Optimalizace:* Metoda zvaná optimalizační postup je podobná předešlým dvěma kromě toho, že dovoluje znovuzařazení objektů. Když se v průběhu tvorby shluků stane, že některý objekt se octne blíže jinému shluku, než ve kterém se právě nachází, optimalizační postup ho přeřadí do jiného bližšího shluku.

Dendrogramy hierarchického shlukování: Analýzou shluků je možné hodnotit jednak podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost znaků analyzovanou pomocí *dendrogramu znaků.*

Dendrogram shluků (vývojový strom) se konstruuje pouze v případě, kdy je k

dispozici matice původních znaků. Dendrogram podobnosti znaků ukazuje rozlišení znaků ve shlucích. Jeho interpretace je snadná: znaky blízko sebe jsou propojeny spojovací úsečkou hodně nízko, mají malou vzdálenost čili značnou vzájemnou podobnost. Znaky propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

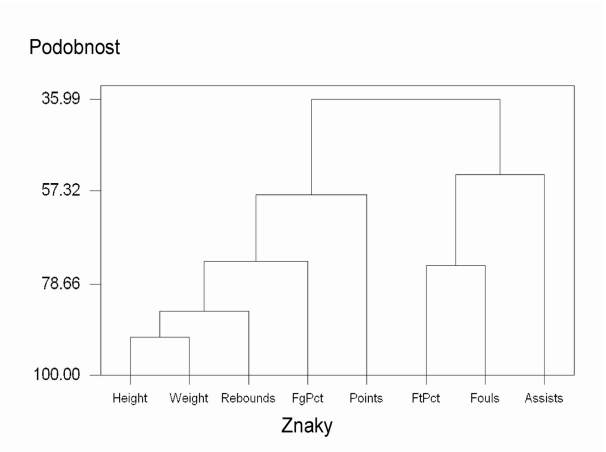
Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku si jsou značně podobné a jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: Dendrogram lze sestrojit celou řadou technik. Prvním kritériem těsnosti proložení při volbě "nejlepšího dendrogramu", jež nejlépe odpovídá struktuře objektů a znaků mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

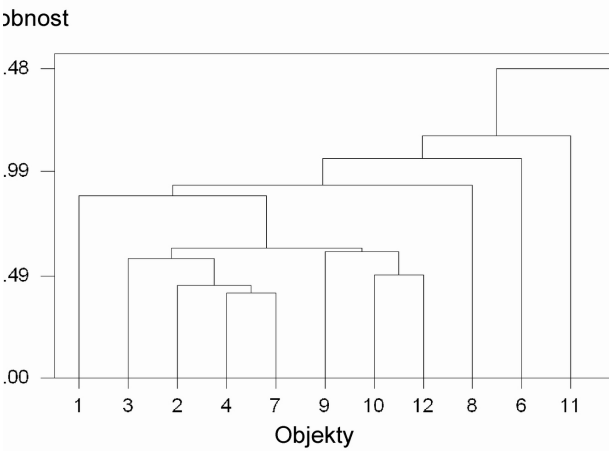
Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^*|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^*)^{1/A}} \right]^A \qquad (12)$$

kde $A = 0.5$ nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu. Je žádoucí, aby hodnoty *delta* byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.



Obr. 4 Dendrogram proměnných metodou průměrné vzdálenosti.



Obr. 5 Dendrogram objektů metodou průměrné vzdálenosti.

Vzorová úloha 1. *Vytvoření dendrogramu objektů neuroleptika*

Neuroleptika redukují nežádoucí účinky přebytečného dopaminu. Liší se však ve svých účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění atd. Účelem je provést klasifikaci neuroleptik do shluků podobných účinků s ohledem na čtyři znaky. Tlumicí účinek neuroleptika se vyjadřuje *mediánovou účinnou dávkou ED50* [mg/kg], která představuje mg léčiva na 1 kg hmotnosti organismu. Častěji se však užívá *převrácená hodnota 1/ED50* [kg/mg], která vyjadřuje přímou úměru mezi množstvím neuroleptika a tlumicí aktivitou. Účelem je provést klasifikaci neuroleptik do shluků podobných účinků.

Data: Charakter proměnných (převrácená hodnota mediánové účinné dávky 1/ED50 [kg/mg]): *B402i* název neuroleptika, *B402x1* potlačení nervozity, *B402x2* potlačení stereotypního chování, *B402x3* potlačení záchvatu a třesu, a *B402x4* dávka smrtícího účinku.

B402*i*	B402*x*2	B402*x*3	B402*x*4	B402*x*5
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

Řešení: Po výběru optimální metody tvorby dendrogramu uvedeme dendrogram podobnosti znaků a dendrogram podobnosti objektů (NCSS2000). Nejvyšší hodnota *kofenetického korelačního koeficientu CC* a nejnižší hodnota kritéria *delta* svědčí o optimální shlukovací metodě tvorby dendrogramu metodou průměrovou.

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;

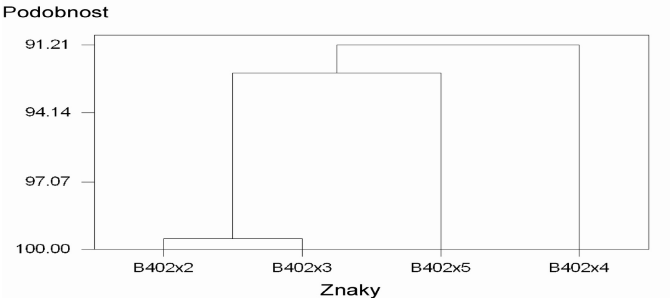
3. Metoda shlukování: Těžiště, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická*

korelace: 0.984750, *Delta(0.5):* 0.175238, *Delta(1.0):* 0.166599;

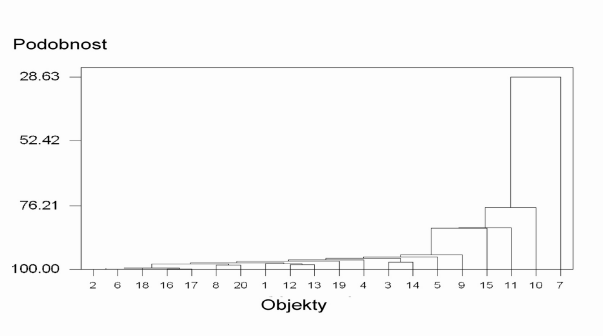
4. Metoda shlukování: Nejbližšího souseda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.988598, *Delta(0.5):* 0.474238, *Delta(1.0):* 0.391993;

5. Metoda shlukování: Median, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.984215, *Delta(0.5):* 0.452308, *Delta(1.0):* 0.428346;

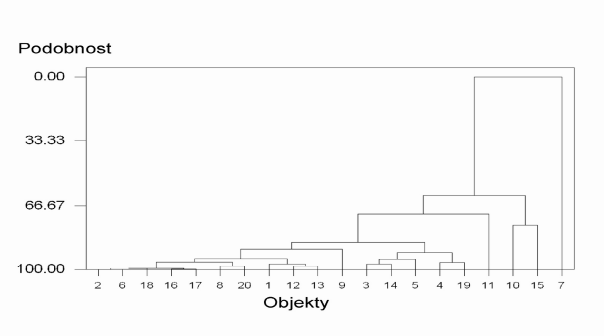
6. Metoda shlukování: Wardova metoda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.979285, *Delta(0.5):* 0.549394, *Delta(1.0):* 0.492716.



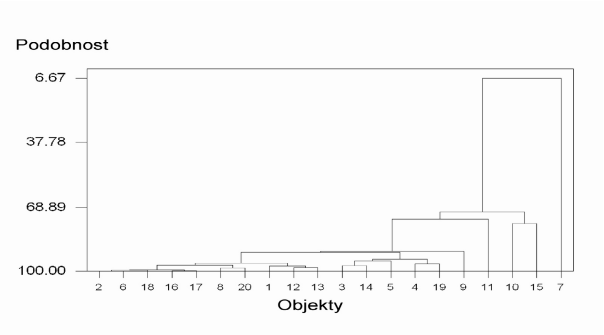
Obr. 6 Metoda průměru v dendrogramu znaků.



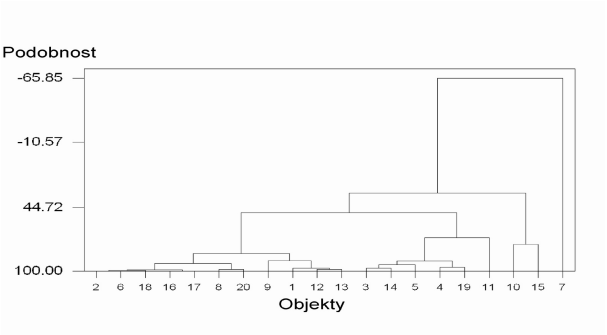
Obr. 7 Dendrogram objektů neuroleptik metodou nejbližšího souseda, *CC* = 0.9886, Δ(0.5) = 0.474, Δ(1.0) = 0.392.



Obr. 8 Dendrogram objektů neuroleptik metodou nejvzdálenějšího souseda, *CC* = 0.9828, Δ(0.5) = 0.179, Δ(1.0) = 0.183.



Obr. 9 Dendrogram objektů neuroleptik metodou průměrné vzdálenosti, *CC* = 0.9889, Δ(0.5) = 0.178, Δ(1.0) = 0.189.



Obr. 10 Dendrogram objektů neuroleptik Wardovou metodou, *CC* = 0.979, Δ(0.5) = 0.549, Δ(1.0) = 0.493.

Metoda skupinového průměru v dendrogramu podobnosti objektů ukazuje na 6 shluků: první shluk obsahuje 10 objektů 18, 16, 17, 8, 20, 1, 12, 13, druhý shluk 5 objektů 3, 14, 5, 4, 19, třetí shluk 2 objekty 10 a 15 a zbývající tři shluky obsahují vždy po jednom objektu, a to 9, 11 a 7. Jako silně vybočující objekt se jeví objekt 7. Dendrogramy získané ostatními technikami dospěly převážně ke stejným shlukům a podobné struktuře.

Závěr: Dle kritérií delta se jako nejvhodnější shlukovací metoda jeví v tomto příkladu dvě metody, a to metoda nejvzdálenějšího souseda a metoda průměrné vzdálenosti.

Shlukování metodou nejbližších těžišť (K-Means)

Metoda nejbližších těžišť poskytuje pouze jediné řešení pro počet požadovaných shluků. Počet shluků musí být zadán uživatelem. Postup je založen na *nejbližším těžišti*, kdy objekt je zařazen do shluku s nejmenší vzdálenosti mezi objektem a těžištěm shluku. Jinak jsou těžiště shluků určována iteračně z dat.

1. Klasifikace, když těžiště shluků jsou známa: Metoda požaduje uživatelem zadaný počet shluků, proto je třeba nejprve aplikovat hierarchickou analýzu shluků na náhodný výběr objektů a určit počet shluků.

2. Klasifikace, když těžiště shluků nejsou známa: Existuje celá řada případů, kdy těžiště shluků nejsou předem známa. Správná těžiště dobře separují shluky. V následujících krocích nahradí objekt těžiště, když jeho nejmenší vzdálenost k těžišti bude větší než vzdálenost mezi dvěma nejbližšími těžišti.

Shlukování metodou optimálních medoidů

Medoid, čili *optimální střed shluku*, je střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Je-li požadováno k shluků, bude existovat také k medoidů. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu. Medoidy a shluky se vytvářejí na základě *vzdáleností* čili *nepodobností*.

Druhů možných znaků v tomto případě shlukování je celá řada: *Intervalové* jsou spojité kladné či záporné, v lineární škále, např. výška, hmotnost, cena, teplota, čas atd. *Ordinální* jsou pořadová čísla stupnice, hodnotící nějakou vlastnost, např. silný nesouhlas (5), nesouhlas (4), neutrální (3), souhlas (2) a silný souhlas (1). *Poměrové* jsou kladné hodnoty, kdy vzdálenost mezi dvěma čísly je stejná, když i jejich poměr je stejný, např. mezi 3 a 30 je stejná jako mezi 30 a 300, chemická koncentrace, intenzita záření, absorbance atd. *Nominální* jsou znaky, které vyjadřují pouze kvalitu a nikoliv kvantitu, např. PSČ, rasa, barva, název města atd. *Symetrické binární:* mají dvě možnosti, obvykle ano (1), ne (0). *Asymetrické binární:* přítomnost či nepřítomnost zřídka se vyskytující události, kdy nepřítomnost není tak důležitá, např. osoba má jizvu na tváři, a tím je lépe identifikovatelná.

Späthova metoda. Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná u počátečního uspořádání shluků, algoritmus pak najde lokální minimum inteligentním přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, proces končí. Lokální minimum však nemusí být globálním. Aby program překonal toto omezení, zopakuje se několikrát hledání vždy z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Jako účelová funkce se bere celková vzdálenost mezi všemi objekty ve shluku podle vzorce

$$D = \sum_{l=1}^{k} \dot{\sum_{i \in c_k}} \dot{\sum_{j \in c_k}} d_{ij} \qquad (13)$$

kde k je celkový počet shluků, d_{ij} je vzdálenost mezi i-tým a j-tým objektem a c_k je soubor všech objektů ve shluku l.

Silueta: Je to statistika, která poskytuje klíčovou informaci o dobrém a špatném shluku. Hodnota siluety *s* se vypočte tímto postupem:

1. Objekt *i* je ve shluku A a má průměrnou vzdálenost *a* ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, je $a = 0$.

2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu *i* ve shluku A a *b* je průměrná vzdálenost mezi objektem *i* a všemi objekty ve shluku B.

3. *Silueta s* objektu *i* se vyčíslí: když shluk A obsahuje pouze jeden objekt, je $s = 0$. Když $a < b$, *je* $s = 1 - a/b$. Když $a > b$, je $s = b/a - 1$. Když $a = b$, je $s = 0$.

Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů nejbližšího souseda B dle pravidla:

1. Je-li *s* blízko hodnotě +1, objekt *i* je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda B.

2. Je-li *s* blízko hodnoty nula, objekt *i* se nachází kdesi uprostřed mezi shluky A a B, a čistě náhodou byl přiřazen do shluku A.

3. Je-li *s* blízko hodnotě -1, objekt *i* je špatně klasifikován. Vzdálenosti k ostatním objektů ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda B. Otázkou pak je, proč byl vlastně zařazen do shluku A.

Určení počtu shluků. Přehlednou statistikou je průměrná silueta *s*, počítaná přes všechny objekty. Tato hodnota sumarizuje jak těsně prokládá shlukové uspořádání analyzovanými daty. Nalezení správného počtu shluků spočívá v nalezení takového počtu, který maximalizuje průměrnou siluetu. Označme *maximální hodnotu průměrné siluety* všech shluků *k* symbolem *SC* a pak budeme rozlišovat následující typy shlukových uspořádání:

SC	Vysvětlení uspořádání do shluků
od 0.71 do 1.00	Silná a dobrá struktura.
od 0.51 do 0.70	Ještě přijatelná struktura.
od 0.26 do 0.50	Slabá struktura, asi umělá. Je třeba najít novou, lepší.
od -1.00 do 0.25	Naprosto nevhodná struktura.

Postup obecné analýzy shluků

Analýza shluků je vždy silnou analytickou pomůckou k účelům zjednodušení, průzkumu a potvrzení struktury.

1. krok: Cíle analýzy shluků: Primárním cílem je rozdělení souboru objektů do dvou nebo více skupin, tříd či shluků. Sledujeme tři cíle:

(a) *Popis systematiky:* Tradičním využitím jsou průzkumové cíle a popis systematiky - *taxonomie*, tj. empirická klasifikace objektů. Analýzou shluků se dospěje k určitým shlukům objektů, které jsou pak porovnány s jejich teoreticky odvozenou typologií.

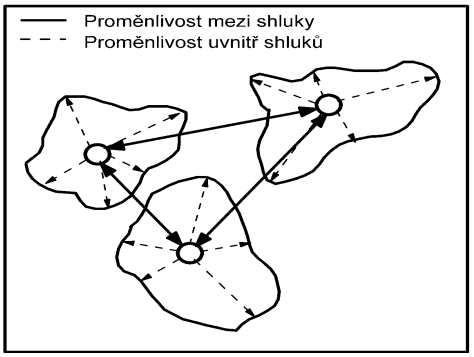
(b) *Zjednodušení dat:* Při hledání taxonomie poskytuje analýza shluků zjednodušený pohled na objekty. Zatímco faktorová analýza se snaží nalézt strukturu znaků, analýza shluků činí totéž ale pro objekty. Na objekty se pak už nehledí jako na jeden společný soubor ale na oddělené shluky objektů, rozčleněné dle jejich vlastností.

(c) *Identifikace vztahu:* Po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty, což by bylo mezi samotnými objekty daleko obtížnější. Shluky mohou být předmětem dalšího kvalitativního uvažování.

2. krok: Formulace úlohy analýzy shluků: S vybranými znaky budeme shlukovat vyšetřované objekty. Nejprve však je třeba odpovědět na tři otázky: (1) Mohou být v datech nějaké odlehlé objekty, které mohou být posléze odstraněny? (2) Jak vyjádříme podobnost objektů? (3) Měla by být data před analýzou shluků standardizována?

3. krok: Předpoklady analýzy shluků: Analýza shluků objektů není charakteru statistického testování. Požadavky normality, linearity, homoskedasticity, které jsou tolik důležité v ostatních vícerozměrných technikách nemají zde význam. Přesto existují dva kritické před-poklady: reprezentativnost a vliv multikolinearity.

Reprezentativnost vzorku: Předpokládá se, že výběr objektů a odvozené shluky reprezentují strukturu celého souboru. Uživatel si proto musí být jist, že zvolený výběr dat je opravdovým představitelem sledovaného souboru.



Obr. 11 Porovnání vzdálenosti mezi shluky a uvnitř shluku.

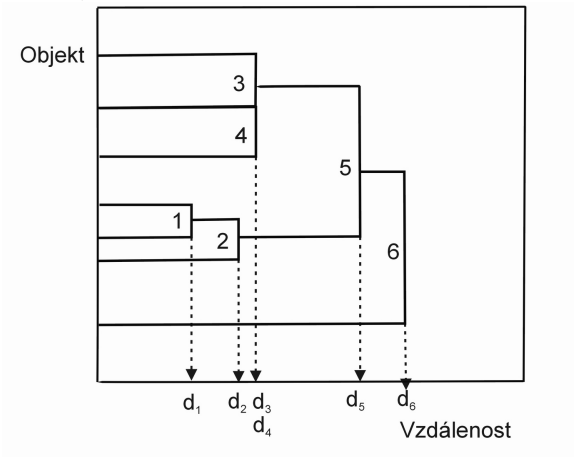
Vliv multikolinearity: Multikolinearita se chová jako neviditelný proces vážení, který silně ovlivňuje analýzu. Uživatel musí proto ověřit přítomnost multikolinearity a když je tato prokázána, je třeba zredukovat počet znaků nebo použít vhodnou míru, jako např. Mahalanobisova vzdálenost, která je schopná multikolinearitu kompenzovat.

4. krok: Výstavba shluků a celková těsnost proložení. Za vhodné rozlišovací kritérium je možné použít maximalizaci rozdílů mezi shluky, a toto

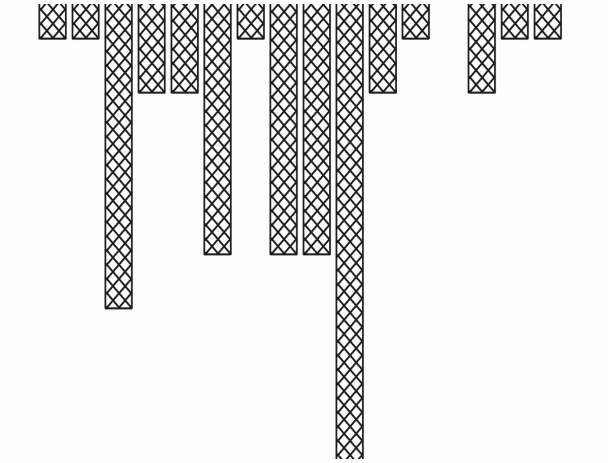
porovnávat vůči proměnli-vosti uvnitř shluků. Testování poměru roztylu mezi jednotlivými shluky vůči průměru rozptylu uvnitř shluku lze porovnat s *F* kritériem v analýze rozptylu.

Hierarchické shlukování: Tyto algoritmy se týkají konstrukce stromovité struktury shluků, dendrogramu. Existují v zásadě dva typy hierarchického shlukování, *aglomerační* a *divizní*. Grafickým zobrazením růstového stromu je diagram, také zvaný *dendrogram*. Jinou grafickou metodou je *vertikální rampouchovitý diagram*.

Když shlukovací proces probíhá v opačném směru než aglomerační, označuje se jako *metoda divizní.* Postup začíná z jednoho velkého shluku, ve kterém jsou všechny objekty. V následujících krocích jsou nepodobné objekty odloučeny ze společného shluku a vzniká tím menší shluk. Proces probíhá tak dlouho, až je ve shluku jediný objekt. Pět nejpoužívanějších aglomeračních algoritmů výstavby shluků jsou metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, metoda průměrová, Wardova metoda, a metoda těžiště.



Obr. 12 Postupná výstavba dendrogramu



Obr. 13 Výstavba rampouchovitého diagramu

Nehierarchické shlukování: Na rozdíl od předešlých hierarchických metod se tyto metody netýkají výstavby stromu, dendrogramu. Místo toho se přidělují objekty do předem známého počtu shluků. Prvním krokem je zadání *zárodečného shluku* jako počátečního středu shluku a všechny objekty nacházející se uvnitř předspecifikované vzdálenosti pak budou do výsledného shluku zařazeny. Pak je zvolen *zárodek* jiného shluku a zařazování do shluku pokračuje až jsou všechny objekty zařazeny.

Přednost hierachickým nebo nehierarchickým metodám? Výzkumný problém musí jedno-značně vést na jednu nebo na druhou metodu. Metodu zvolíme dle obsahu úlohy.

(1) *Důvody pro a proti hierarchickým metodám:* V minulosti byly hierarchické shlukové techniky populárnější. Wardova a průměrová metoda byly považovány za nejlepší techniky. Hierarchické metody mají výhodu, že jsou rychlé a spotřebují méně strojového času.

(2) *Užívání nehierarchických metod:* Nehierachické metody se v poslední doběvyužívají stále více. Jejich kvalita závisí na schopnosti uživatele, jeho praktických zkušenostech a objektivní teorii jak si vybrat zárodkové body.

(3) *Kombinace obou metod:* Jiným přístupem je užití *obou* metod, hierarchické a nehierarchické, abychom využili výhod obou. Nejprve hierarchickou metodou určíme počet shluků, profily shlukovaných center a identifikují se zřetelné odlehlé body.

Počet vytvářených shluků: Snad nejvíce matoucí otázkou v analýze shluků je dosažení konečného počtu shluků, známého také pod názvem *terminační kritérium*. Neexistuje žádný objektivní způsob určení tohoto kritéria. Jedno z terminačních kritérií se týká relativně jednoduchého vyšetření měr podobnosti mezi shluky v každém kroku, když totiž míra podobnosti překročí předdefinovanou velikost nebo když následné hodnoty se skokově změní. Je vhodné postupovat tak, že se určí rozličný počet shluků např. 2, 3 a 4 a na základě úvah o alternativním řešení, praktickém úsudku a teoretických základech úlohy samé se rozhodne. Když se objeví jednoobjektový shluk nebo shluk o poměrně malé velikosti, uživatel musí rozhodnout, zda tento představuje strukturního člena vzorku nebo zda ho lze označit jako nedostatečně reprezentativní pro soubor dat.

5. krok: Interpretace shluků: Interpretace shluků se týká vyšetření každého shluku v pojmech shlukových znaků a především pojmenování shluků, které vystihuje podstatu a povahu shluků. Při zahájení interpretace si uvědomíme, že ve shlukové analýze je často užívanou mírou těžiště shluků. Uživatel se *musí* vrátit k původním datům v původních znacích a vyčíslit průměrové profily pro původní data. Vyšetříme proto profily průměrových skóre a označíme popisným nadpisem každý shluk.

6. krok: Validace a profilování shluků: Existuje poněkud subjektivní charakter analýzy shluků stran hledání optimálního shlukového řešení.

Validace shluků: Validace čili potvrzení shluků zahrnuje pokusy zajistit, že shlukové řešení je reprezentativní v celé obecné populaci, a je proto zobecnitelné i na ostatní objekty a dále je i stabilní v čase.

Profilování shlukového řešení: Profilování se týká popisu vlastností každého shluku, aby se objasnilo jak se vlastnosti liší ve významných dimenzích. Profilová analýza se zaměřuje na popis ne toho, co konkrétně odhalují shluky ale na vlastnosti shluků, na jejímž základě byly identifikovány. Kromě toho je kladen důraz na vlastnosti, ve kterých se shluky vzájemně odlišují a na vlastnosti, které mohou predikovat účast v dotyčném shluku.

Ilustrace shlukové analýzy:

Vzorová úloha 2. *Popis a třídění polétavých mšic* (B404)

Jeffers (1967) studoval 40 jedinců polétavých mšic *(Alate adelges)* za pomoci světelné pasti. U mšic bylo změřeno 19 ukazatelů, sloužících k rozlišení druhů a typů tohoto hmyzu: 14 proměnných se týká délky nebo šířky, 4 proměnné se týkají počtu a 1 proměnná je binární, vyjadřující přítomnost či absenci. Mšice se totiž obtížně rozlišují dle běžných taxonometrických klíčů a obvykle je nutné detailní vyšetření dat, hledání společných rysů a následná klasifikace do větších celků.

Data: *B404i* index jedince, *B404x*1 délka těla, *B404x*2 šířka těla, *B404x*3 délka předního křídla, *B404x*4 délka zadního křídla, *B404x*5 počet průduchů, *B404x*6 délka tykadla I, *B404x*7 délka tykadla II, *B404x*8 délka tykadla III, *B404x*9 délka tykadla IV, *B404x*10 délka tykadla V, *B404x*11 počet tykadlových ostnů, *B404x*12 délka posledního článku

nohy, *B404x*13 délka holeně, tibia, *B404x*14 délka stehna, *B404x*15 délka sosáku, *B404x*16 délka kladélka, *B404x*17 počet kladélkových trnů, *B404x*18 řitní otvor, *B404x*19 počet háčků zadních křídel. (Data jsou k dispozici na internetu na adrese: //meloun.upce.cz/data a dále pak Kompendium-Data).

i	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}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8	3.3	3	4.4	4.5	3.6	7.0	4.0	8	0	3.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
40	12.8	5.7	4.8	2.8	5.0	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5.0	3.1	8	1	2.0

Vzorová úloha 3. *Klasifikace zdrojů pitné vody* (E404)

Na 62 vzorcích pitné vody, u kterých bylo provedeno stanovení 16 proměnných kvality, proveďte vícerozměrnou statistickou analýzu.

Data: E404i index vzorku, *E404x1* obsah dusičnanů [mg/l], *E404x2* obsah dusitanů [mg/l], *E404x3* obsah chloridů [mg/l], *E404x4* obsah celkového chloru [mg/l], *E404x5* obsah síranů [mg/l], *E404x6* obsah fosforečnanů [mg/l], *E404x7* obsah amonných solí [mg/l], *E404x8* obsah vápníku [mg/l], *E404x9* obsah hořčíku [mg/l], *E404x10* obsah železa (celkového) [mg/l], *E404x11* obsah manganu [mg/l], *E404x12* pH, *E404x13* KNK, *E404x14* ZNK, *E404x15* vodivost, *E404x16* nerozpuštěné látky [mg/l]. (Data jsou k dispozici na internetu na adrese: //meloun.upce.cz/data a dále pak Kompendium-Data).

i	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}	x_{14}	x_{15}	x_{16}
1	2.2	0.00	6.	6.	103.5	0.032	0.02	181	17	0.016	0.05	7.08	8.1	3.40	855	0.09
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
62	32.8	0.01	25.	25.	115.5	0.050	0.02	102	12	0.016	0.05	7.69	2.6	0.65	436	0.05

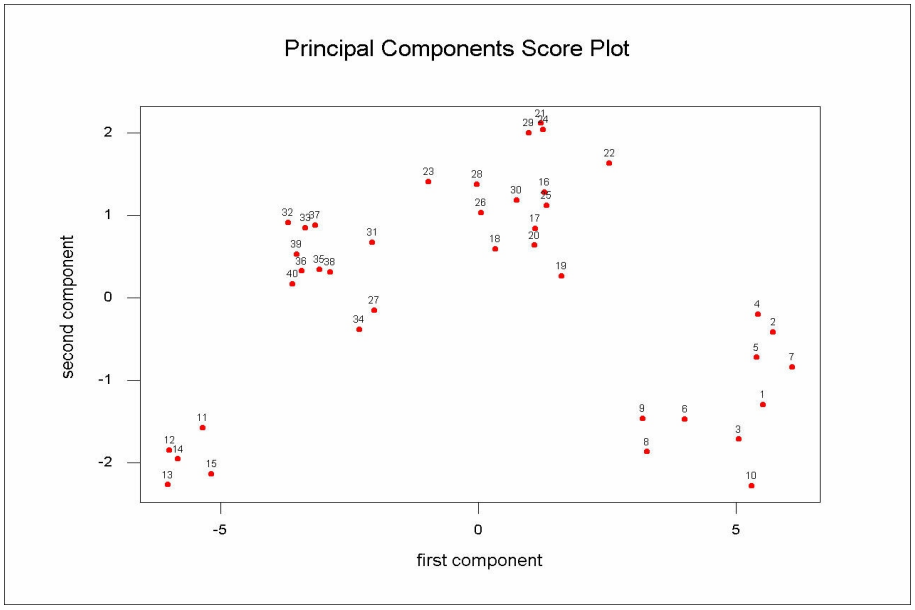
Vzorová úloha 4. *Klasifikace vlastností rozličných druhů kávy* (E406)

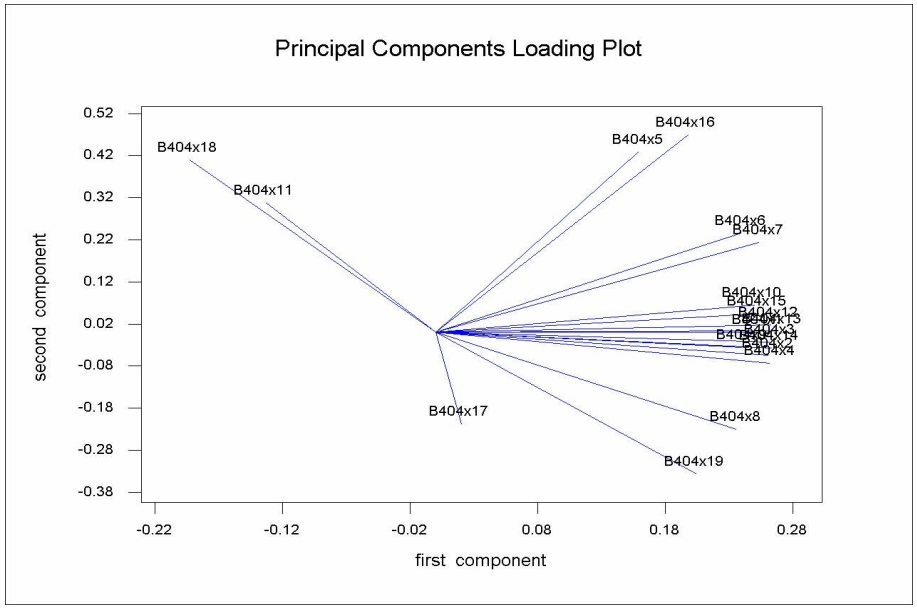
Byl získán výběr 43 vzorků kávy, pocházejících ze 30 zemí. U každého druhu kávy byly změřeny jeho chemické a fyzikální vlastnosti. Testujte tato data, zda-li splňují požadavky na homogenitu, nebo zda je možné indikovat dvě či více rozličných kategorií.

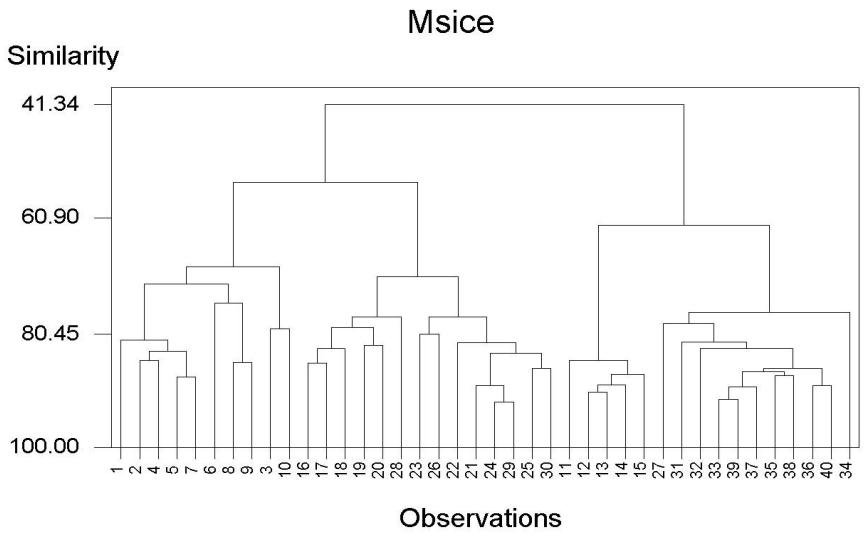
Data: Soubor dat COFFEE obsahuje na 43 vzorcích kávy (řádky) popis pomocí 13 proměnných (sloupce) v pořadí: *E406i* index kávy, *E406j* je původ kávy, *E406x1* obsah vody, *E406x2* hmotnost zrn, *E406x3* extrakt, *E406x4* pH, *E406x5* volná acidita, *E406x6* obsah minerálů, *E406x7* tuky, *E406x8* kofein, *E406x9* trinonelin, *E406x10* kyselina chlorogeniková, *E406x11* kyselina neochlorogeniková, *E406x12* kyseliny isochlorogeniková, *E406x13* suma kyselin chlorogenikových. (Data jsou k dispozici na internetu na adrese: //meloun.upce.cz/data a dále pak Kompendium-Data).

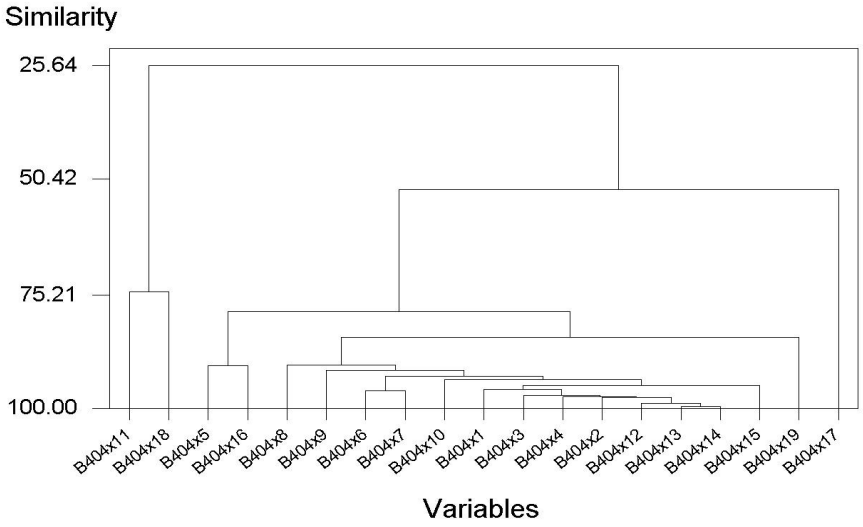
i	j	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1.0	5.4	0.4	0.8	6.6
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5

Řešení vzorové úlohy 2.

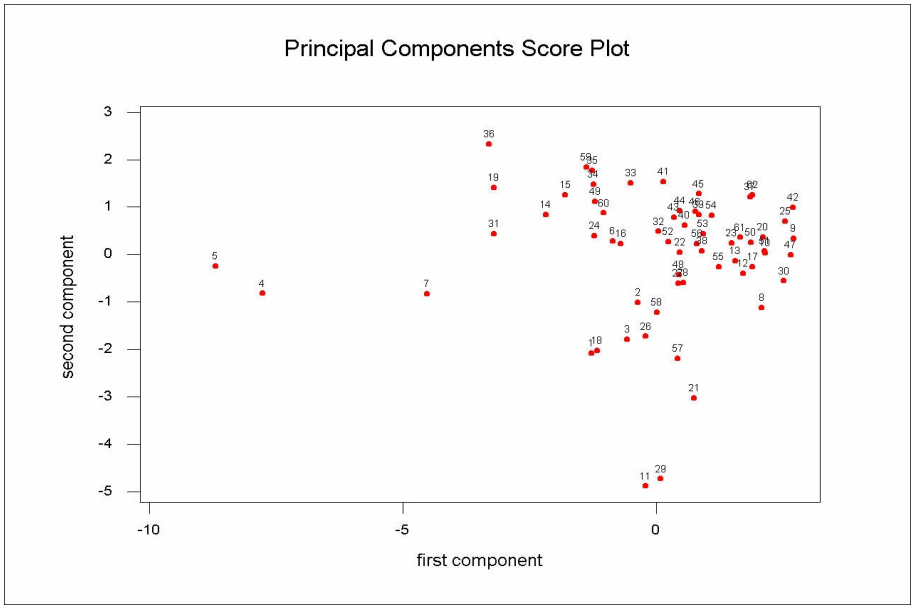


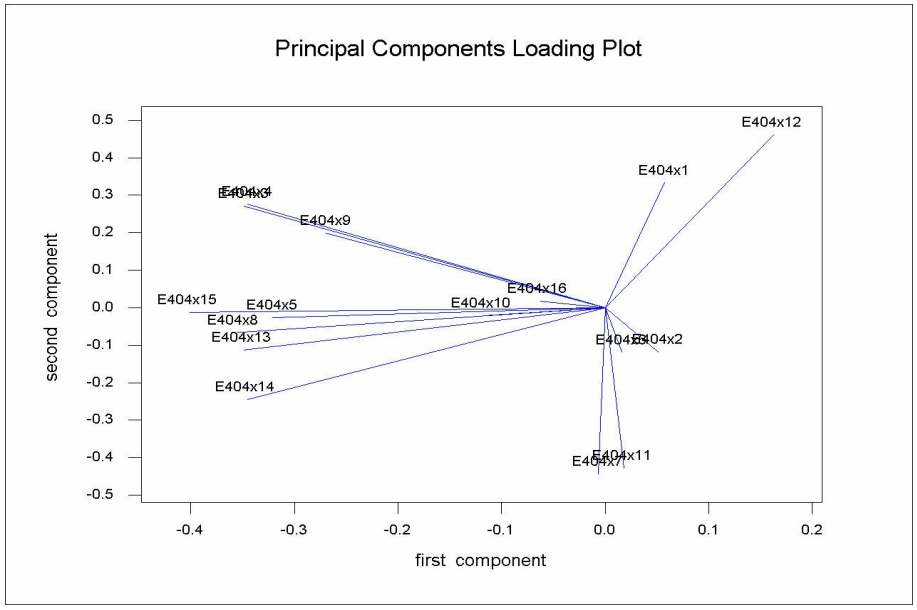


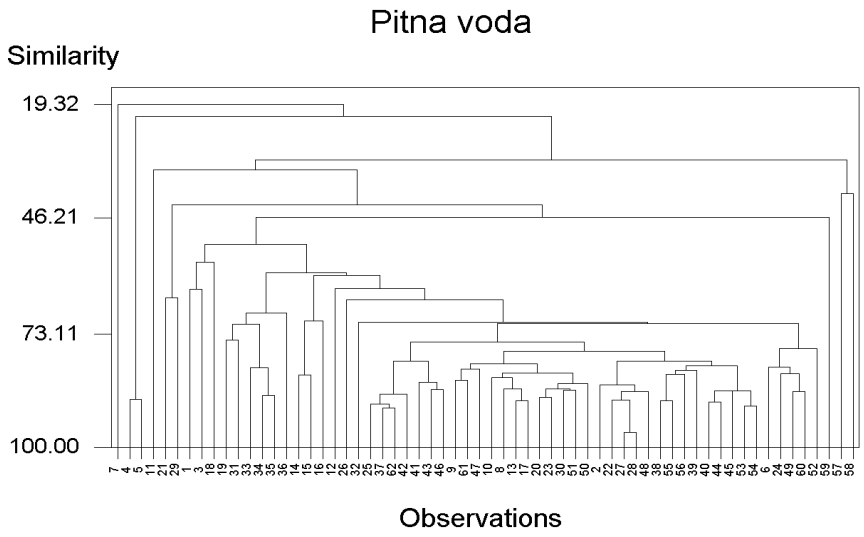


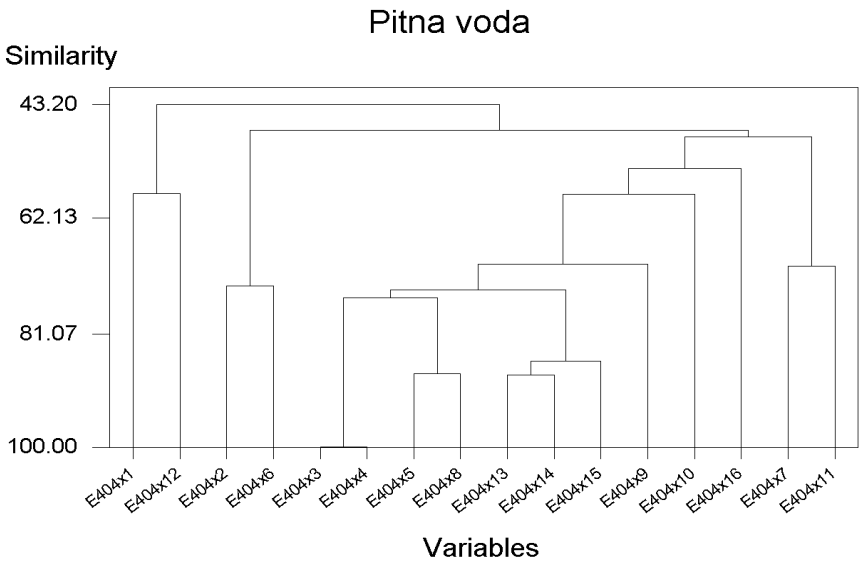
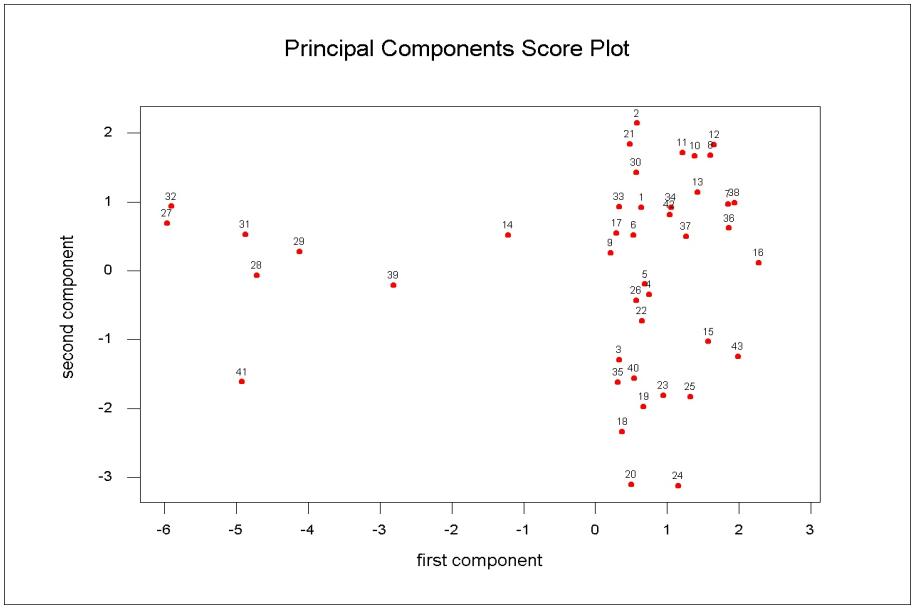
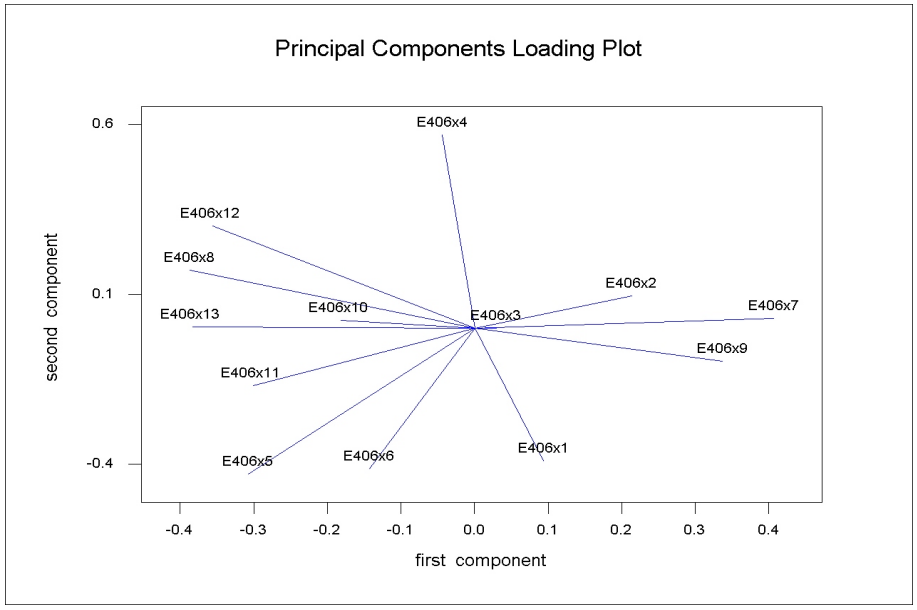
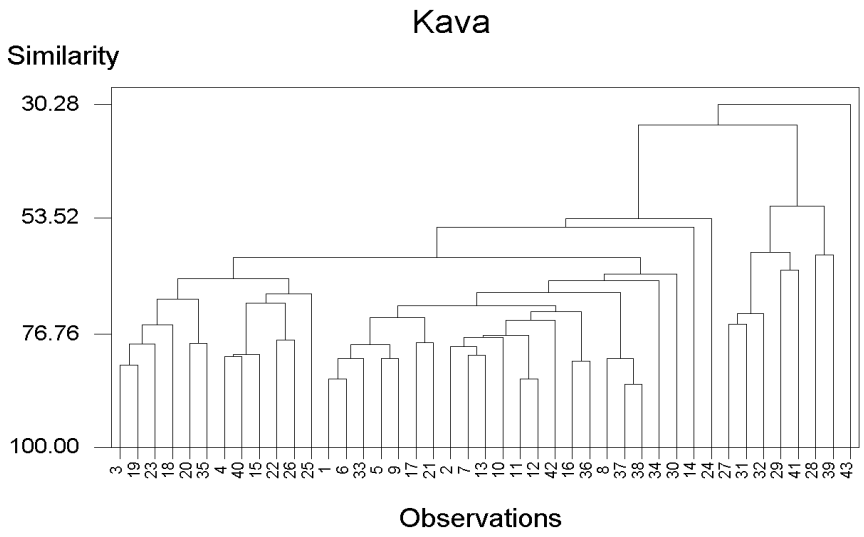
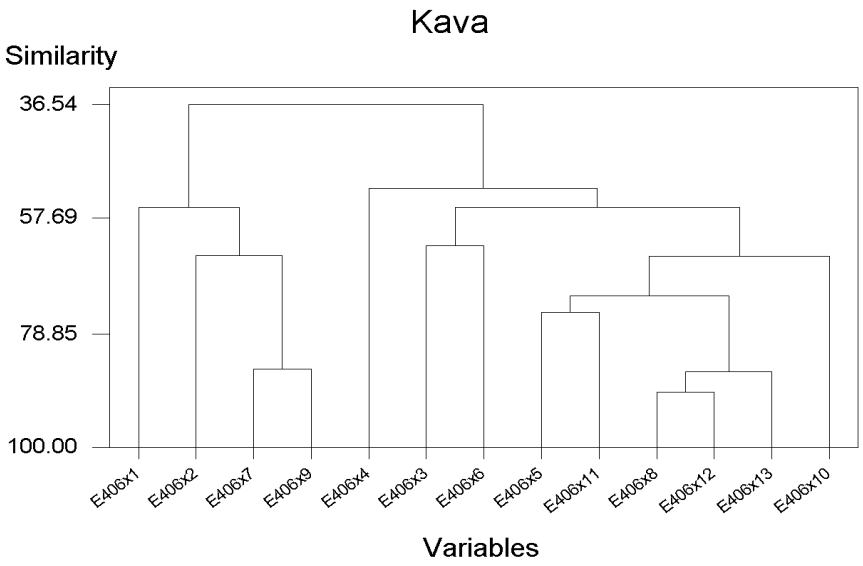


Řešení vzorové úlohy 3.









Řešení vzorové úlohy 4.