4. Průzkumová analýza vícerozměrných dat

4.1. Motivace

Při statistickém zpracování dat se často setkáváme s vícerozměrnými daty. Vyskytují se v situacích, kdy u každého z n objektů zjišťujeme hodnoty p znaků $X_1, ..., X_p$. Dostáváme tak p-rozměrný datový soubor ve formě matice n x p:

$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$. Řádky této matice se vztahují k jednotlivým objektům, zatímco sloupce k jednotlivým znakům.

Průzkumová analýza vícerozměrných dat je daleko složitější než analýza jednorozměrných dat a zahrnuje velké množství úloh, které směřují např. k pochopení závislostí mezi znaky. V této kapitole se omezíme na dva problémy, a to na vizualizaci dat pomocí hlavních komponent a na shlukovou analýzu dat.

4.2. Vizualizace vícerozměrných dat

V případě dvourozměrného či třírozměrného datového souboru lze vlastnosti objektu popsané znaky graficky znázornit tak, že hodnoty znaků chápeme jako souřadnice v dvourozměrném či třírozměrném prostoru a získáme tak dvourozměrný či třírozměrný tečkový diagram. Z jeho vzhledu lze poznat, zda se v datech vyskytují odlehlá pozorování, zda mezi znaky existuje nějaká závislost nebo zda se objekty sdružují do skupin.

Je-li $p > 3$, použijeme k vizualizaci dat metodu hlavních komponent, která informace o variabilitě obsažené v datovém souboru dokáže vyjádřit pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních. Data pak znázorníme v prostoru prvních dvou či tří hlavních komponent.

4.3. Podstata metody hlavních komponent

4.3.1. Označení

$\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ – vektor pozorování i-tého objektu, i = 1, 2, ..., n

$m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ - průměr j-tého znaku, j = 1, 2, ..., p

$s_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - m_j\right)^2$ - rozptyl j-tého znaku, j = 1, 2, ..., p

$z_{ij} = \frac{x_{ij} - m_j}{s_j}$ - (i,j)-tá standardizovaná hodnota, i = 1, 2, ..., n, j = 1, 2, ..., p

$\mathbf{z}_i = (z_{i1}, ..., z_{ip})^T$ – vektor standardizovaných pozorování i-tého objektu, i = 1, 2, ..., n

$\mathbf{m} = (m_1, ..., m_p)^T$ – vektor průměrů

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \mathbf{m}\right)\left(\mathbf{x}_i - \mathbf{m}\right)^T$ - výběrová varianční matice

$\mathbf{R} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^T$ - výběrová korelační matice

(**S** a **R** jsou čtvercové symetrické matice řádu p.)

4.3.2. Základní pojmy

Nechť **A** je čtvercová matice řádu p. Číslo λ se nazývá vlastní číslo matice **A**, jestliže pro libovolný nenulový vektor **v** typu p x 1 splňuje rovnici $\mathbf{Av} = \lambda\mathbf{v}$. Vektor **v** se nazývá vlastní vektor matice **A**. Determinant $|\mathbf{A} - \lambda\mathbf{I}|$ se nazývá charakteristický polynom matice **A**. Stopou matice **A** rozumíme součet jejích diagonálních prvků.

Rovnici $\mathbf{Av} = \lambda\mathbf{v}$ upravíme na tvar $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{o}$. Tato soustava p rovnic má netriviální řešení, právě když charakteristický polynom matice **A** je roven 0. Dostaneme rovnici p-tého stupně. Jejím řešením jsou vlastní čísla $\lambda_1, \ldots, \lambda_p$. Jejich součet je roven stopě matice **A**.

4.3.3. Získání hlavních komponent

Nechť výběrová varianční matice **S** má vlastní čísla $l_1, \ldots, l_p$ a vlastní vektory $\mathbf{v}_1, \ldots, \mathbf{v}_p$, přičemž $\mathbf{v}_j^T\mathbf{v}_j = 1$, j = 1, ..., p a $\mathbf{v}_j^T\mathbf{v}_k = 0$ pro $j \neq k$. Znamená to, že vektory $\mathbf{v}_1, \ldots, \mathbf{v}_p$ jsou ortonormální. Bez újmy na obecnosti předpokládáme, že $l_1 > l_2 > \ldots > l_p$.

1. hlavní komponenta Y_1 vznikne jako lineární kombinace znaků $X_1, \ldots, X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_1$, tedy
$Y_1 = v_{11}X_1 + \ldots + v_{1p}X_p$. Její rozptyl je l_1.

Dosadíme-li za $X_1, \ldots, X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n})^T$, kde $y_{1i} = \mathbf{v}_1^T\mathbf{x}_i$.

2. hlavní komponenta vznikne jako lineární kombinace znaků $X_1, \ldots, X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_2$, tedy

$Y_2 = v_{21}X_1 + \ldots + v_{2p}X_p$. Její rozptyl je l_2. Přitom $\mathbf{v}_1^T\mathbf{v}_2 = 0$, tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.

Dosadíme-li za $X_1, \ldots, X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_2 = (y_{21}, \ldots, y_{2n})^T$, kde $y_{2i} = \mathbf{v}_2^T\mathbf{x}_i$.

.....

j-tá hlavní komponenta vznikne jako lineární kombinace znaků $X_1, \ldots, X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_j$, tedy

$Y_j = v_{j1}X_1 + \ldots + v_{jp}X_p$. Její rozptyl je l_j. Přitom $\mathbf{v}_j^T\mathbf{v}_k = 0$, j = 1, ..., k-1, tj. j-tá hlavní komponenta je lineárně nezávislá se všemi ostatními hlavními komponentami.

Dosadíme-li za $X_1, \ldots, X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})^T$, kde $y_{ji} = \mathbf{v}_j^T\mathbf{x}_i$.

Lze dokázat, že celková variabilita obsažená v datech je rovna stopě matice **S**, tj. součtu vlastních čísel $l_1 + \ldots + l_p$. 1. hlavní komponenta tedy vyčerpává $\frac{l_1}{l_1 + \ldots + l_p}100\%$ celkové variability. Pokud je číslo $\frac{l_1}{l_1 + \ldots + l_p}$ dostatečně blízké 1, znamená to, že 1. hlavní komponenta dobře nahrazuje celý datový soubor. Je-li toto číslo podstatně menší než 1, musíme vzít tolik hlavních komponent, aby jejich součet dělený stopou matice **S** byl dostatečně blízký 1. V mnoha aplikacích se stává, že i při velkém počtu znaků stačí poměrně malý počet hlavních komponent.

Znázorníme-li rozmístění objektů na ploše prvních dvou hlavních komponent, můžeme poznat, které objekty se řadí do skupin neboli shluků.

(Před provedením metody hlavních komponent je třeba se rozhodnout, zda budeme pracovat s původními hodnotami znaků nebo standardizovanými hodnotami.)

4.4. Shluková analýza

4.4.1. Cíl shlukové analýzy

Cílem shlukové analýzy je roztřídění n objektů, z nichž každý je popsán p znaky, do několika pokud možno stejnorodých (homogenních) skupin (shluků, clusterů). Požadujeme, aby objekty uvnitř shluků si byly podobné co nejvíce, zatímco objekty z různých shluků co nejméně. Přesný počet shluků většinou není přesně znám.

Shluková analýza nachází uplatnění v celé řadě oborů, např. v biologii. U n populací změříme p biometrických charakteristik a zjišťujeme, zda určité skupiny populací tvoří shluky.

Shluková analýza je ovšem průzkumovou metodou a měla by sloužit jako určité vodítko při dalším zpracování dat.

4.4.2. Podobnost objektů

Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost. Nechť k-tý objekt je popsán vektorem pozorování $\mathbf{x}_k = (x_{k1}, ..., x_{kp})^T$ a l-tý objekt vektorem $\mathbf{x}_l = (x_{l1}, ..., x_{lp})^T$. Euklidovská vzdálenost k-tého a l-tého objektu: $d_{kl} = \sqrt{\sum_{j=1}^{p} \left(x_{kj} - x_{lj}\right)^2}$.

Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

4.4.3. Hierarchické shlukování

Při aplikacích shlukové analýzy se nejčastěji používá aglomerativní hierarchická procedura. Její princip spočívá v postupném slučování objektů, a to nejprve nejbližších a v dalších krocích pak stále vzdálenějších.

Algoritmus:

1. krok: Každý objekt považujeme za samostatný shluk.
2. krok: Najdeme dva shluky, jejichž vzdálenost je minimální.
3. krok: Tyto dva shluky spojíme v nový, větší shluk a přepočítáme matici vzdáleností. Její řád se sníží o 1. Vrátíme se na 2. krok.

Funkce algoritmu končí, až jsou všechny objekty spojeny do jediného shluku.

Vzdálenost mezi shluky se počítá různými způsoby. Uvedeme tři z nich.

a) Metoda nejbližšího souseda: Vzdálenost mezi dvěma shluky je minimem ze všech vzdáleností mezi jejich objekty.

b) Metoda nejvzdálenějšího souseda: Vzdálenost mezi dvěma shluky je maximem ze všech vzdáleností mezi jejich objekty.

c) Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.

Výsledky aglomerativní hierarchické procedury se zpravidla znázorňují pomocí dendrogramu. Je to graficky znázorněná posloupnost dvojic $\left\{\left(\nu_1, S^{(1)}\right), \ldots, \left(\nu_n, S^{(n)}\right)\right\}$, kde $\{\nu_i\}_{i=1}^{n}$ je neklesající posloupnost úrovní spojování a $S^{(i)}$ je roztřídění objektů odpovídající úrovni ν_i, $i = 1, ..., n$.

4.4.4. Příklad

Pro 5 objektů byla vypočtena matice euklidovských vzdáleností. Proveďte shlukovou analýzu všemi třemi uvedenými metodami a výsledky znázorněte pomocí dendrogramů.

$$\begin{pmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}$$

Řešení pro metodu nejbližšího souseda

$d_{35} = 2$, tedy shluky 3 a 5 vytvoří shluk (35). Zjistíme vzdálenost mezi shlukem (35) a shluky 1, 2, 4.

$d_{(35)1} = \min(d_{31}, d_{51}) = \min(3, 11) = 3$, $d_{(35)2} = \min(d_{32}, d_{52}) = \min(7, 10) = 7$,

$d_{(35)4} = \min(d_{34}, d_{54}) = \min(9, 8) = 8$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 3 & 7 & 8 \\ 3 & 0 & 9 & 6 \\ 7 & 9 & 0 & 5 \\ 8 & 6 & 5 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{(35)1} = 3$, tedy shluk 1 a shluk (35) vytvoří shluk (135). Zjistíme vzdálenost mezi shlukem (135) a shluky 2 a 4.

$d_{(135)2} = \min(d_{12}, d_{(35)2}) = \min(9, 7) = 7$, $d_{(135)4} = \min(d_{14}, d_{(35)4}) = \min(6, 8) = 6$.

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 7 & 6 \\ 7 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{24} = 5$, tedy shluk 2 a shluk 4 vytvoří shluk (24). Zjistíme vzdálenost mezi shlukem (135) a shlukem (24).

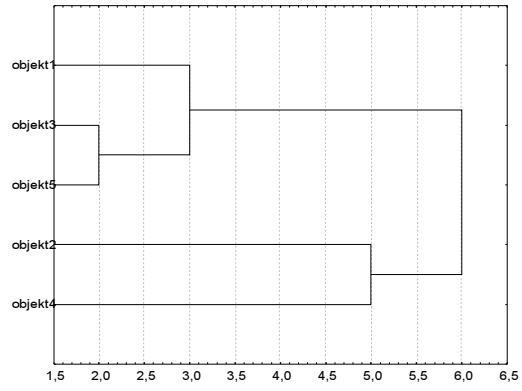
$d_{(135)(24)} = \min(d_{(135)2}, d_{(135)4}) = \min(7, 6) = 6$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 6 \\ 6 & 0 \end{pmatrix}$$

Nakonec shluky (135) a (24) vytvoří jediný shluk(12345).

Dendrogram



Řešení pro metodu nejvzdálenějšího souseda

$d_{35} = 2$, tedy shluky 3 a 5 vytvoří shluk (35). Zjistíme vzdálenost mezi shlukem (35) a shluky 1, 2, 4.

$d_{(35)1} = \max(d_{31}, d_{51}) = \max(3, 11) = 11$, $d_{(35)2} = \max(d_{32}, d_{52}) = \max(7, 10) = 10$,

$d_{(35)4} = \max(d_{34}, d_{54}) = \max(9, 8) = 9$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 11 & 10 & 9 \\ 11 & 0 & 9 & 6 \\ 10 & 9 & 0 & 5 \\ 9 & 6 & 5 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{24} = 5$, tedy objekt 2 a objekt 4 vytvoří shluk (24). Zjistíme vzdálenost mezi shlukem (35) a shlukem (24) a mezi shlukem (24) a objektem 1.

$d_{(35)(24)} = \max(d_{(35)2}, d_{(35)4}) = \max(10, 9) = 10$, $d_{(24)1} = \max(d_{21}, d_{41}) = \max(9, 6) = 9$.

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 10 & 11 \\ 10 & 0 & 9 \\ 11 & 9 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{(24)1} = 9$ tedy shluk (24) a objekt 1 vytvoří shluk (124). Zjistíme vzdálenost mezi shlukem (35) a shlukem (124).

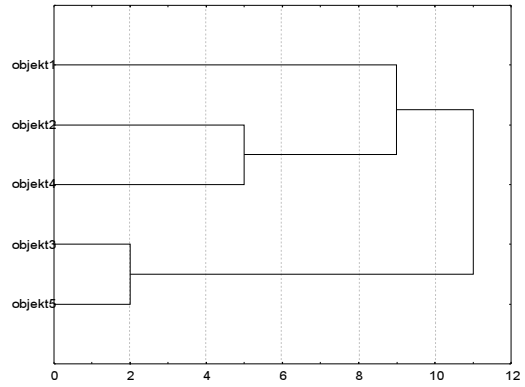
$d_{(35)(124)} = \max(d_{(35)(24)}, d_{(35)1}) = \max(10, 11) = 11$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 11 \\ 11 & 0 \end{pmatrix}$$

Nakonec shluky (35) a (124) vytvoří jediný shluk(12345).

Dendrogram



Řešení pro metodu průměrné vazby

$d_{35} = 2$, tedy shluky 3 a 5 vytvoří shluk (35). Zjistíme vzdálenost mezi shlukem (35) a shluky 1, 2, 4.

$d_{(35)1} = 0{,}5(d_{31} + d_{51}) = 0{,}5(3 + 11) = 7$, $d_{(35)2} = 0{,}5(d_{32} + d_{52}) = 0{,}5(7 + 10) = 8{,}5$,

$d_{(35)4} = 0{,}5(d_{34} + d_{54}) = 0{,}5(9 + 8) = 8{,}5$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 7 & 8{,}5 & 8{,}5 \\ 7 & 0 & 9 & 6 \\ 8{,}5 & 9 & 0 & 5 \\ 8{,}5 & 6 & 5 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{24} = 5$, tedy objekt 2 a objekt 4 vytvoří shluk (24). Zjistíme vzdálenost mezi shlukem (35) a shlukem (24) a mezi shlukem (24) a objektem 1.

$d_{(35)(24)} = 0{,}5(d_{(35)2} + d_{(35)4}) = 0{,}5(8{,}5 + 8{,}5) = 8{,}5$, $d_{(24)1} = 0{,}5(d_{21} + d_{41}) = 0{,}5(9 + 6) = 7{,}5$.

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 8{,}5 & 7 \\ 8{,}5 & 0 & 7{,}5 \\ 7 & 7{,}5 & 0 \end{pmatrix}$$

V této matici najdeme nejmenší číslo: $d_{(35)1} = 7$ tedy shluk (35) a objekt 1 vytvoří shluk (135). Zjistíme vzdálenost mezi shlukem (135) a shlukem (24).

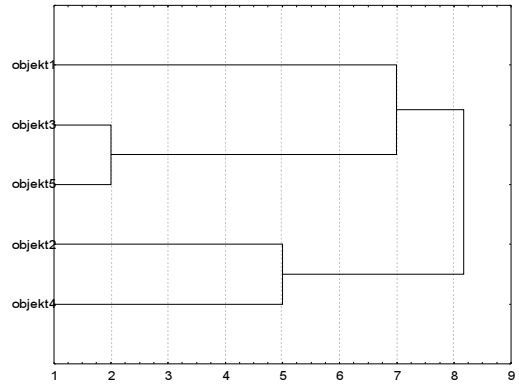
$d_{(24)(135)} = 0{,}5(d_{(24)1}, d_{(24)(35)}) = 0{,}5(7{,}5 + 8{,}5) = 8$

Sestavíme novou matici vzdáleností

$$\begin{pmatrix} 0 & 8 \\ 8 & 0 \end{pmatrix}$$

Nakonec shluky (135) a (24) vytvoří jediný shluk(12345).

Dendrogram



Příklady ke 4. kapitole

Příklad 1.: Na pěti objektech byly zjišťovány hodnoty dvou znaků. Datový soubor je tvaru $\begin{pmatrix} 3 & 7 \\ 5 & 6 \\ 6 & 8 \\ 7 & 10 \\ 9 & 9 \end{pmatrix}$. Najděte vektor průměrů, výběrovou varianční matici a výběrovou korelační matici.

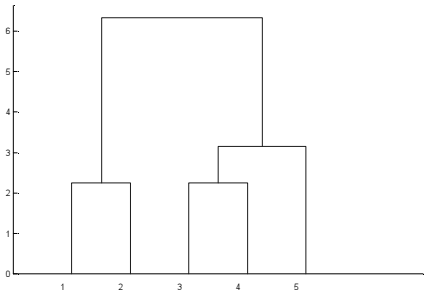
(**m** = $(6,8)^T$, **S** = $\begin{pmatrix} 5 & 2,5 \\ 2,5 & 2,5 \end{pmatrix}$, **R** = $\begin{pmatrix} 1 & 0,707 \\ 0,707 & 1 \end{pmatrix}$)

Příklad 2.:Na 24 objektech byly pozorovány znaky X_1, X_2 a X_3. Z datového souboru byla vypočtena výběrová varianční matice **S** = $\begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 169,70 & 103,29 & 66,69 \end{pmatrix}$. Vlastní čísla získaná řešením rovnice $|\mathbf{S} - l\mathbf{I}| = 0$ a jim odpovídající vlastní vektory jsou: $l_1 = 680,411$, $l_2 = 6,5016$, $l_3 = 2,8573$, $\mathbf{v}_1 = (0,8126; 0,4955; 0,3068)^T$, $\mathbf{v}_2 = (0,5454; -0,8321; -0,1009)^T$, $\mathbf{v}_3 = (0,2053; 0,2493; -0,9464)^T$. Vyjádřete hlavní komponenty a určete, kolik procent variability obsažené v matici **S** každá z nich vyčerpává.
($Y_1 = 0,8126\ X_1 + 0,4955\ X_2 + 0,3068\ X_3$, vyčerpává 98,65% celkové variability,
$Y_2 = 0,5454\ X_1 - 0,8321\ X_2 - 0,1009\ X_3$, vyčerpává 0,94% celkové variability,
$Y_3 = 0,2053\ X_1 + 0,2493\ X_2 - 0,9464\ X_3$, vyčerpává 0,41% celkové variability.)

Příklad 3.: Pro údaje z příkladu 1. sestavte matici euklidovských vzdáleností.

$$\begin{pmatrix} 0 & \sqrt{5} & \sqrt{10} & 5 & 2\sqrt{10} \\ \sqrt{5} & 0 & \sqrt{5} & 2\sqrt{5} & 5 \\ \sqrt{10} & 5 & 0 & \sqrt{5} & \sqrt{10} \\ 5 & 2\sqrt{5} & \sqrt{5} & 0 & \sqrt{5} \\ 2\sqrt{10} & 5 & \sqrt{10} & \sqrt{5} & 0 \end{pmatrix}$$

Příklad 4.: Pomocí matice euklidovských vzdáleností vypočtené v příkladě 3. proveďte shlukovou analýzu metodou nejvzdálenějšího souseda. Výsledek znázorněte dendrogramem



Práce se systémem STATISTICA

Téma: Shluková analýza

V souboru stanice.sta jsou uloženy údaje (v $\mu g/m^3$) o průměrných ročních koncentracích oxidu siřičitého v letech 1993 – 1998 na deseti brněnských měřicích stanicích: Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice, Tuřany. Cílem je najít metodami shlukové analýzy skupiny stanic, které vykazují podobné rysy chování.

Úkoly:

1. Soubor stanice.sta upravte tak, aby případy 1 až 10 byly pojmenovány názvy stanic.
   Návod: Data – Case Name Manager – Case name length 3, Transfer case names From Variable Stanice, OK.
2. Prozkoumejte proměnné r93 až r98 pomocí krabicových diagramů.
   Návod: Graphs – 2D Graphs – Box Plots – Graph Type Multiple – Variables r93, ..., r98, OK, OK.
3. Z krabicových diagramů je vidět, že proměnné r93 až r98 vykazují velmi rozdílnou variabilitu. Vytvořte proto standardizované proměnné a nadále pracujte s nimi.
   Návod: Data – Standardize – Variables r93, ..., r98, OK.
4. Z proměnných r93 až r98 vytvořte dvě hlavní komponenty.
   Návod: Statsitics – Multivariate Exploratory Techniques – Principal Component & Classification Analysis / Variables r93, ..., r98, OK, OK – Number of factors 2, zaškrtněte Plot case factor coordinates 2D.
   Interpretace: Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.
5. Pro standardizované proměnné r93 až r98 proveďte shlukovou analýzu s euklidovskou vzdáleností a třemi metodami: nejbližšího souseda, nejvzdálenějšího souseda a průměrné vazby. Výsledky znázorněte pomocí dendrogramu.
   Návod: Statistics - Multivariate Exploratory Techniques – Cluster Analysis – OK- Variables r93, ..., r98, OK, Cluster Cases(row) – OK – Horizontal hierarchical tree plot. Euklidovská vzdálenost a metoda nejbližšího souseda je nastavena implicitně. Pro další dvě metody změňte Amalgamation (linkage) rule ze Single Linkage na Complete resp Unweighted pair-group average.
6. Vypočtěte a pomocí sloupkových diagramů znázorněte průměrné roční koncentrace SO_2 a směrodatné odchylky za celé sledované období pro všech 10 stanic.
   Návod: Je nutné se vrátit k původním nestandardizovaným hodnotám, tj. znovu načíst soubor stanice.sta a pojmenovat případy názvy stanic – viz bod 1. Pak je zapotřebí soubor transponovat – zaměnit řádky za sloupce: Data – Transpose – File. Vymažete 1. řádek: Edit – Delete – Cases – From case 1 To case 1, OK. Pomocí Descriptive Statistics vypočtěte průměry a směrodatné odchylky proměnných DOB až TUR. Vytvoření sloupkových diagramů pro průměry: ve workbooku klikněte pravým tlačítkem myši na sloupek Mean: Graphs of Block Data – Custom Graph for Entire Column – Select Graph Type Bar/Colum Plot, OK. Podobně pro směrodatné odchylky.
   Interpretace: Stanice v 1. shluku (DOB, HUS, KRA, SKA) vykazují za sledované období poměrně nízké průměrné koncentrace SO_2 (od 6 $\mu g/m^3$ po 11 $\mu g/m^3$) i malé směrodatné odchylky (od 2,5 $\mu g/m^3$ po 3,5 $\mu g/m^3$). Druhý shluk obsahuje stanice s vysokými koncentracemi (od 13 $\mu g/m^3$ po 19 $\mu g/m^3$) a velkými směrodatnými odchylkami (od 3,8 $\mu g/m^3$ po 6,8 $\mu g/m^3$.