Přednáška č. 4: Shluková analýza

Motivace:

S problematikou klasifikace objektů do skupin se v praxi setkáváme velmi často. Např. biolog stduduje vnitrodruhovou variabilitu určitého druhu. Na 50 lokálních populacích změří biometrické charakteristiky (jako je délka nejvyššího listu, délka korunní trubky, počet květů apod.) a zjišťuje, zda jsou si určité skupiny populací podobnější než jiné, zda tvoří shluky.

Jako první použil pojem „shluková analýza“ Američan Robert C. Tryon v roce 1939: „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobností a rozdílností.“

Metody hledání shluků můžeme rozdělit na dvě velké skupiny: hierarchické metody a nehierarchické metody.

a) Hierarchické metody vytvářejí shluky, které mají různou hierarchickou úroveň – shluky vyšší hierarchické úrovně obsahují shluky nižší úrovně. Hierarchické metody jsou buď aglomerativní (menší shluky se postupně spojují do větších shluků) nebo divizní (celý soubor je nejprve chápán jako jeden shluk a postupně se dělí na menší shluky. Zde se seznámíme s aglomerativním hierarchickým algoritmem. Výsledky hierarchických metod se graficky znázorňují pomocí dendrogramu, což je binární strom znázorněný buď vertikálně nebo horizontálně. V dendrogramu každý uzel představuje shluk.
   V horizontálním dendrogramu horizontální směr reprezentuje vzdálenosti mezi shluky. Vertikální řezy dendrogramem představují roztřídění objektů do shluků.
b) Nehierarchické metody nevytvářejí hierarchickou strukturu. Rozkládají původní množinu objektů do několika disjunktních shluků tak, aby bylo splněno určité kritérium. Zde se seznámíme s metodou k-průměrů, která umožňuje provést rozklad množiny objektů do předem specifikovaného počtu shluků.

Shluková analýza nachází uplatnění v celé řadě oborů, např. v biologii, psychologii, geografii, technice i marketingu.

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

Cíl shlukové analýzy

Vycházíme z p-rozměrného datového souboru $\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$, který získáme tak, že na každém z n objektů změříme hodnoty p znaků $X_1, \dots, X_p$. Cílem shlukové analýzy je roztřídění těchto n objektů 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 předem znám.

Podobnost objektů

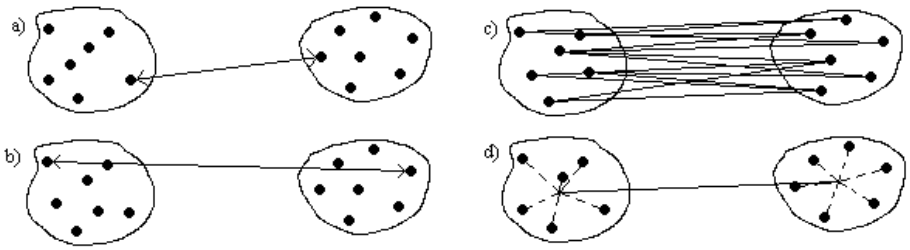
Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost. Nechť k-tý objekt je popsán vektorem pozorování $\mathbf{x}_k = (x_{k1}, ..., x_{kp})^T$ a l-tý objekt vektorem
$\mathbf{x}_l = (x_{l1}, ..., x_{lp})^T$. Euklidovská vzdálenost k-tého a l-tého objektu:
$d_{kl} = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^2}$. Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností $\mathbf{D} = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

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.
d) Wardova metoda: Vybírá takové shluky ke sloučení, kde je minimální součet čtverců odchylek všech pozorování od příslušných shlukových průměrů. Obecně lze říci, že je tato metoda velmi účinná, ale má tendenci tvořit poměrně malé shluky. Požaduje vyjádření vzdálenosti objektů čtvercovou euklidovskou vzdáleností.



Schematické znázornění: a) metoda nejbližšího souseda, b) metoda nejvzdálenějšího souseda, c) metoda průměrné vazby, d) Wardova metoda

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

Kofenetický koeficient korelace

Různé shlukovací procedury mohou poskytovat různé výsledky. K posouzení shody mezi maticí vzdáleností objektů a výsledkem dané shlukovací metody je možno použít např. kofenetický koeficient korelace. Posuzuje míru shody mezi maticí vzdáleností objektů a výsledkem dané shlukovací metody. Je to koeficient korelace mezi n(n-1)/2 prvky umístěnými nad (nebo pod) hlavní diagonálou matice vzdáleností a odpovídajícími prvky kofenetické matice. Přitom (i,j)-tý prvek této matice je definován jako ta vzdálenost i-tého a j-tého objektu, při níž jsou tyto objekty poprvé spojeny do jednoho shluku. Této vzdálenosti se říká kofenetická vzdálenost. Z uvažovaných shlukovacích metod pak vybereme tu, která poskytuje nejvyšší kofenetický koeficient korelace.
Upozornění: Systémy STATISTICA a SPSS bohužel neposkytují kofenetický koeficient korelace. Je možno ho získat pomocí systému MATLAB.
Návod: Do matice X uložíme zkoumaný datový soubor.
Y = pdist(X, 'euclid') … poskytne řádkový vektor obsahující prvky nad hlavní diagonálou matice euklidovských vzdáleností.
Z = linkage(Y,'single') … poskytne matici o n-1 řádcích a 3 sloupcích, která obsahuje informace potřebné pro sestrojení dendrogramu (parametr single je pro metodu nejbližšího souseda, pro metodu nejvzdálenějšího souseda je complete, pro metodu průměrné vazby average a pro Wardovu metodu ward).
c = cophenet(Z,Y) … poskytne kofenetický koeficient korelace.
dendrogram(Z) … vykreslí se dendrogram pro výsledky zvolené hierarchické aglomerativní procedury.

Příklad: Tento příklad vychází z publikace

Budíková, Marie. Aplikace shlukové analýzy v ekologii. Praha : Jednota českých matematiků a fyziků, 2001. 8 s. Sborník prací 11. letní školy ROBUST 2000.

V rámci jedné z bakalářských prací obhájených na katedře geografie byly shromážděny údaje o průměrných měsíčních koncentracích oxidu siřičitého v letech 1984 – 1998 na 10 monitorovacích stanicích umístěných na území města Brna. Jednalo se o stanice umístěné v lokalitách Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice a Tuřany, ve zkratkách DOB, HUS, KRA, KRO, MZL, POL, PRI, SKA, SOB a TUR. Tyto údaje měly – mimo jiné – posloužit také k řešení problému optimalizace sítě stanic.

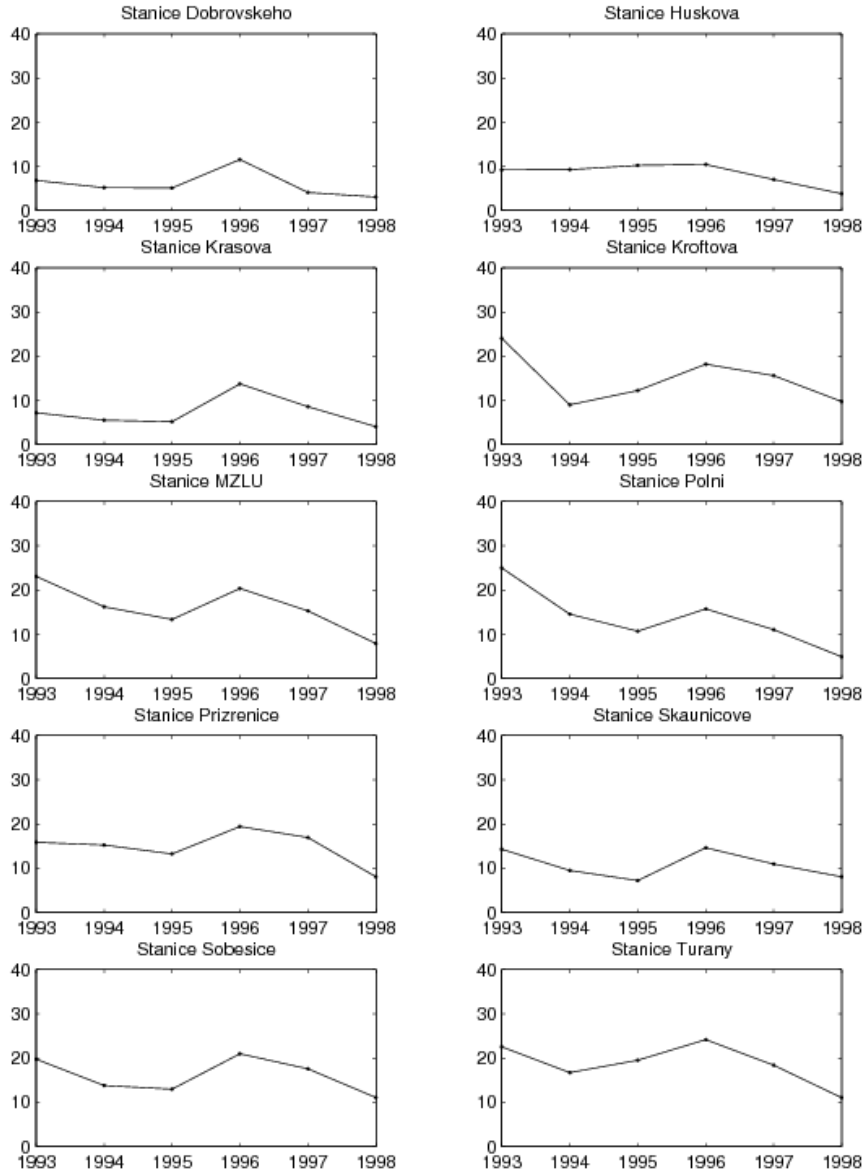
Uvedené stanice jsou obhospodařovány jednak brněnskou pobočkou ČHMÚ (to jsou stanice KRO, MZL, PRI, SOB, TUR) a jednak MHS (to jsou stanice DOB, HUS, KRA, POL, SKA). Každá z těchto organizací však zjišťuje hodnoty SO_2 jinou metodou – ČHMÚ gravimetrickou a MHS aspiračně kolorimetrickou. Teprve od r.1993 jsou výsledky kolorimetrické metody přepočítávány tak, aby odpovídaly výsledkům metody gravimetrické.

Do našeho zpracování byly tedy zahrnuty údaje až od r. 1993, konkrétně jsme se zabývali průměrnými ročními koncentracemi SO_2. Jenom na okraj uvádím, že podle zákona o ochraně ovzduší před znečišťujícími látkami činí nejvyšší přípustná průměrná roční koncentrace SO_2 60 mikrogramů na metr krychlový.

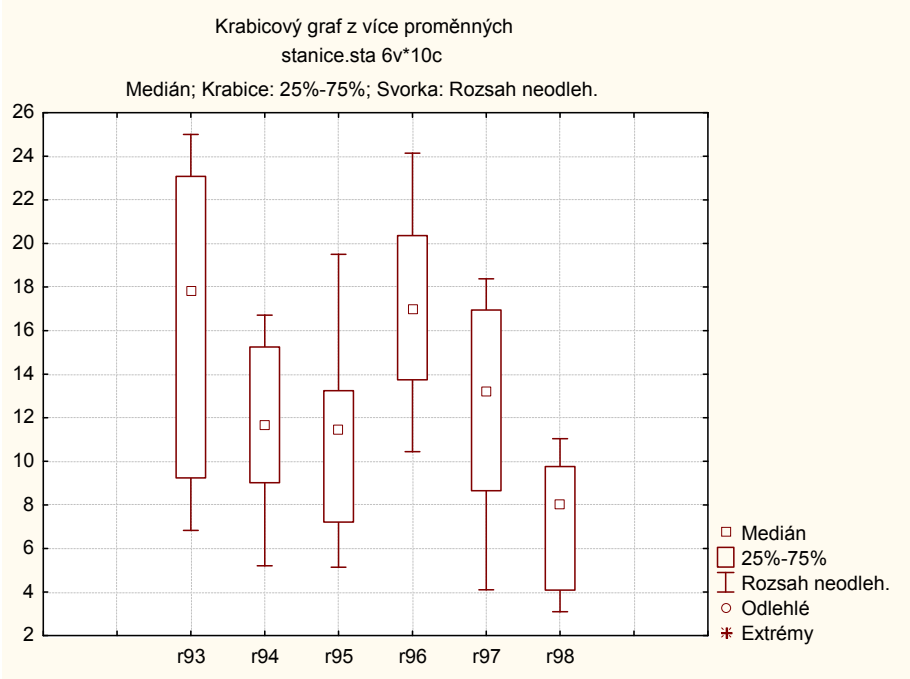
Každá ze sledovaných 10 stanic byly popsána šesti údaji, jak vidíme v této tabulce.

	1 r93	2 r94	3 r95	4 r96	5 r97	6 r98
DOB	6,828	5,202	5,137	11,568	4,104	3,097
HUS	9,241	9,281	10,259	10,442	7,035	3,857
KRA	7,205	5,535	5,197	13,741	8,651	4,085
KRO	24,039	9,018	12,237	18,189	15,601	9,762
MZL	23,079	16,222	13,353	20,363	15,312	7,925
POL	25,005	14,568	10,723	15,76	11,068	4,916
PRI	15,874	15,251	13,241	19,435	16,943	8,081
SKA	14,297	9,49	7,209	14,434	10,961	8,063
SOB	19,728	13,772	12,943	20,948	17,564	11,039
TUR	22,524	16,708	19,502	24,144	18,377	11,024

Časové řady ročních hodnot znečištění na sledovaných stanicích máme znázorněny na následujícím obrázku.



Naším cílem bylo najít stanice, které mají podobné rysy chování, tedy vytvořit skupiny (shluky) takových stanic. Prvním krokem bylo provedení průzkumové analýzy dat pomocí krabicových diagramů.



Na první pohled je zřejmé, že údaje v jednotlivých letech vykazují dosti rozdílnou variabilitu, největší v r. 1993, nejmenší v r. 1998. Provedli jsme tedy standardizaci a nadále pracovali se standardizovanými hodnotami.

Datový soubor standardizovaných hodnot

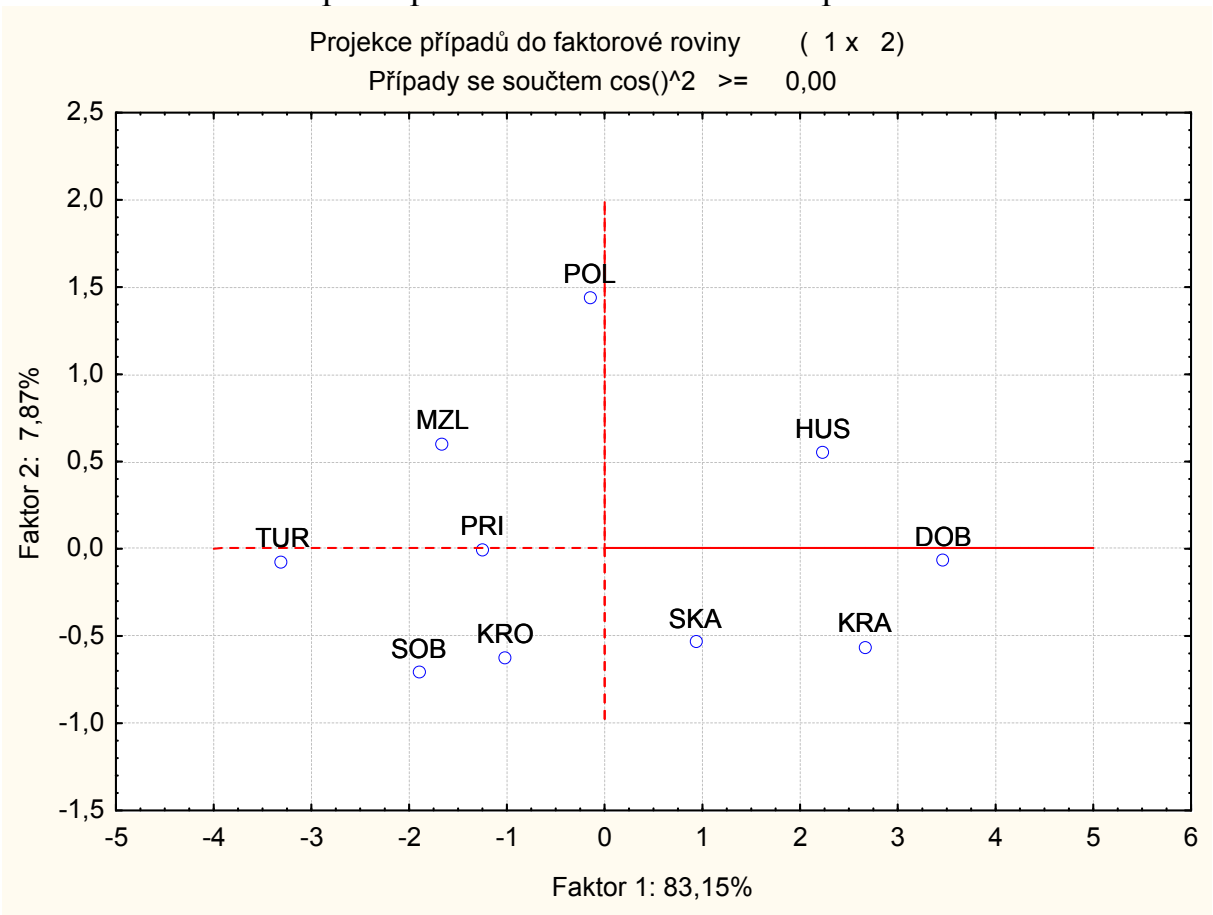
	1 r93	2 r94	3 r95	4 r96	5 r97	6 r98
DOB	-1,398	-1,457	-1,34	-1,205	-1,722	-1,363
HUS	-1,059	-0,514	-0,165	-1,459	-1,126	-1,11
KRA	-1,345	-1,38	-1,326	-0,714	-0,796	-1,034
KRO	1,0192	-0,575	0,2882	0,2906	0,619	0,8596
MZL	0,8844	1,0904	0,5441	0,7816	0,5601	0,2469
POL	1,1549	0,7081	-0,059	-0,258	-0,304	-0,757
PRI	-0,128	0,866	0,5184	0,572	0,8923	0,2989
SKA	-0,349	-0,466	-0,865	-0,557	-0,326	0,2929
SOB	0,4138	0,5241	0,4501	0,9137	1,0188	1,2855
TUR	0,8065	1,2028	1,954	1,6355	1,1843	1,2805

Nyní přistoupíme k vizualizaci dat na ploše prvních dvou hlavních komponent.

Pořadí vl.č.	Vlastní čísla korelační matice a související statistiky (stanice.sta) Pouze aktiv. proměnné			
	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	4,989279	83,15465	4,989279	83,1546
2	0,472272	7,87121	5,461551	91,0259
3	0,300851	5,01419	5,762402	96,0400
4	0,129928	2,16547	5,892330	98,2055
5	0,073190	1,21984	5,965521	99,4253
6	0,034479	0,57466	6,000000	100,0000

1. hlavní komponenta vyčerpává 83,15% variability dat a druhá 7,87%.

Rozmístění stanic na ploše prvních dvou hlavních komponent:

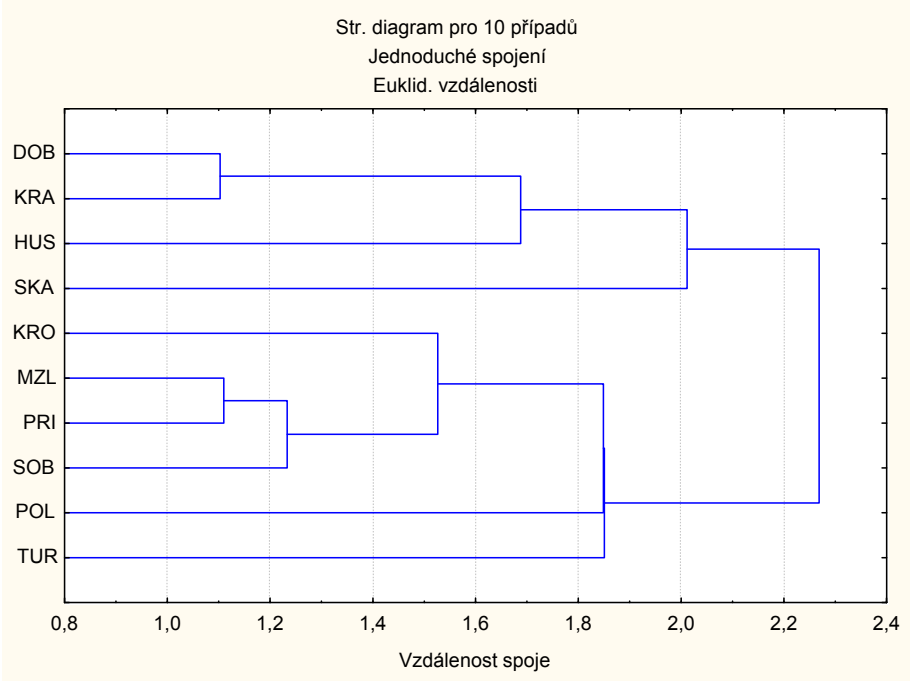


Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

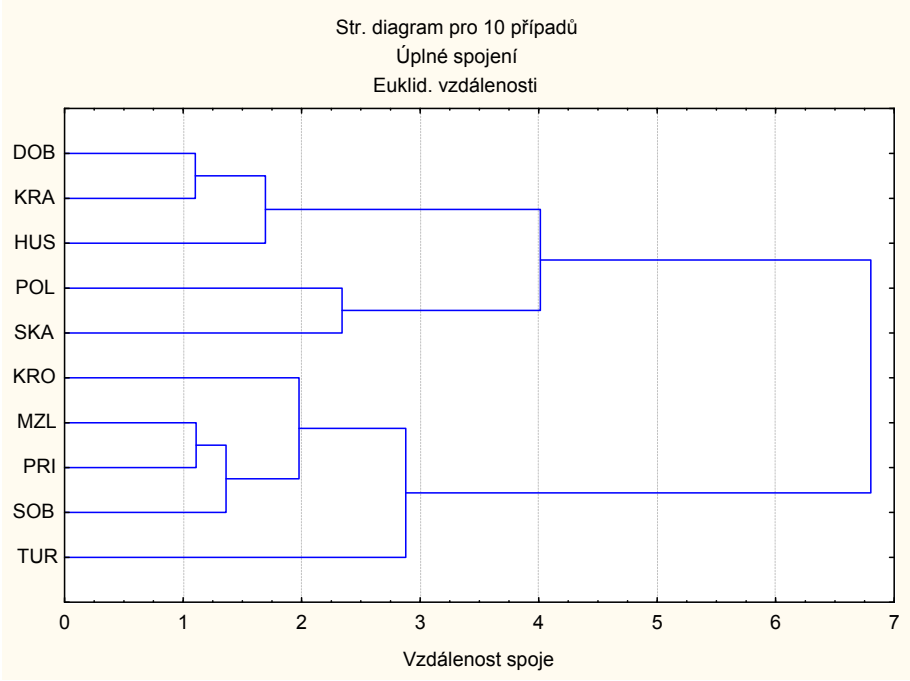
Pro standardizované proměnné r93 až r98 provedeme shlukovou analýzu s euklidovskou vzdáleností a čtyřmi metodami: nejbližšího souseda, nejvzdálenějšího souseda, průměrné vazby a Wardovu metodu. Výsledky znázorníme pomocí dendrogramu.
Návod: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza Spojování (hierarchické shlukování) – OK - Proměnné r93, ..., r98, OK, Detaily -

Shlukovat případy (řádky) – Pravidlo slučování: Jednoduché spojení – Míry vzdálenosti: Euclidovské vzdálenosti - OK – Horizontální graf hierarch. stromu. Euklidovská vzdálenost a metoda nejbližšího souseda je nastavena implicitně. Pro další dvě metody změňte Pravidlo slučování z Jednoduchého spojení na Úplné spojení resp. Nevážený průměr skupin dvojic resp. Wardova metoda.

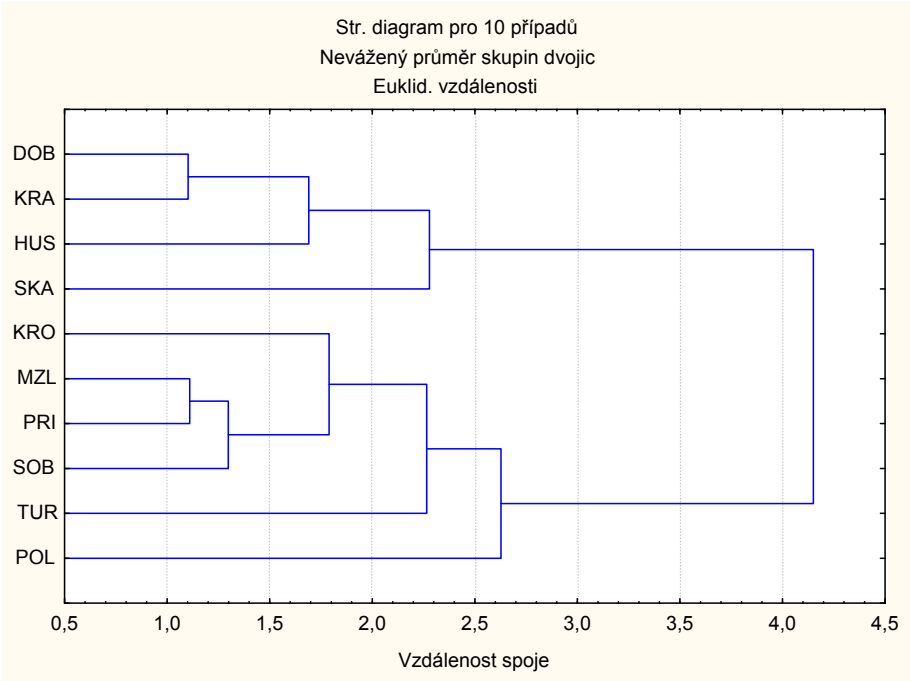
Dendrogram pro metodu nejbližšího souseda:



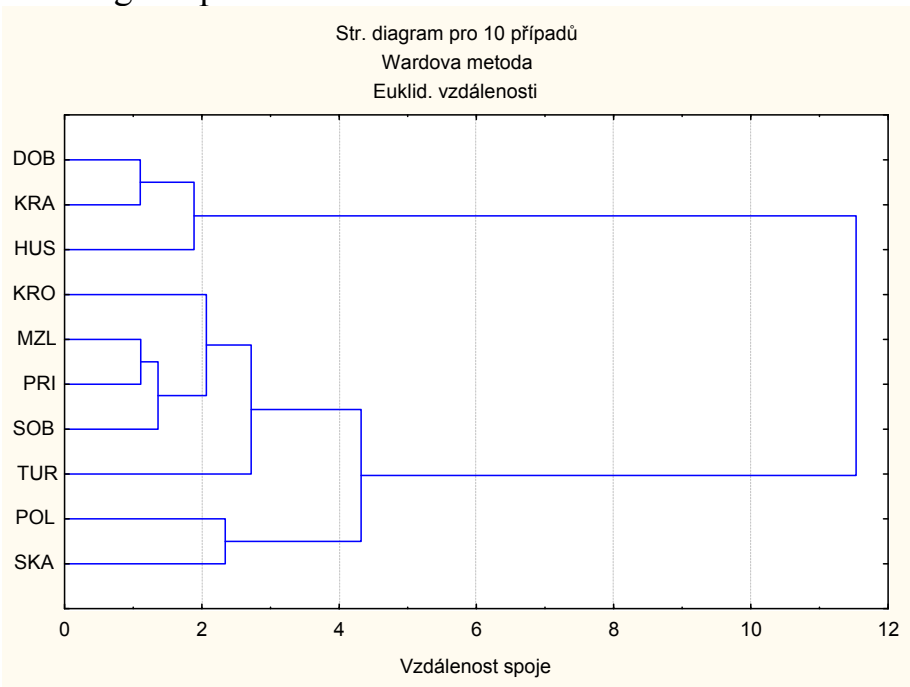
Dendrogram pro metodu nejvzdálenějšího souseda:



Dendrogram pro metodu průměrné vazby:



Dendrogram pro Wardovu metodu:



Uvedené metody dávají poněkud rozdílné výsledky. Shodu mezi maticí vzdáleností a dendrogramem posoudíme pomocí kofenetických koeficientů korelace. Tyto koeficienty byly vypočítány pomocí systému MATLAB.

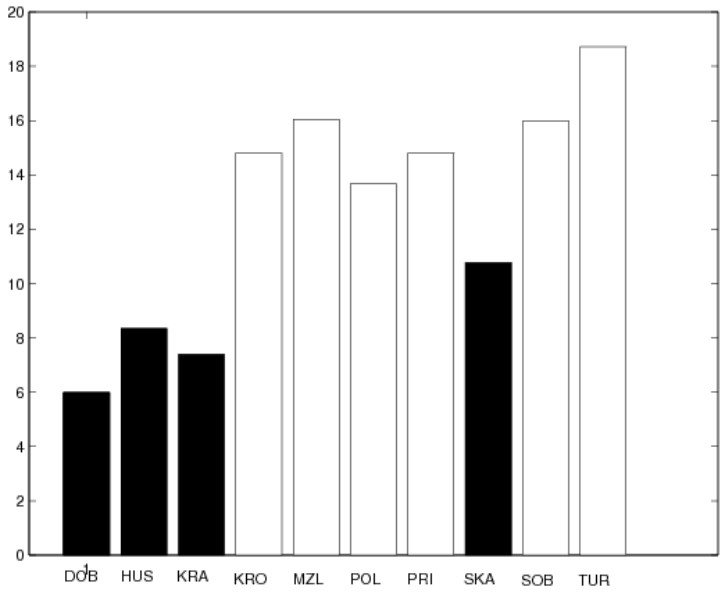
metoda	koefenetický koeficient
nejbližšího souseda	0,8133
nejvzdálenějšího souseda	0,8262
průměrné vazby	0,8312
Wardova	0,8253

Nejvyšší kofenetický koeficient poskytla metoda průměrné vazby, tedy nadále budeme uvažovat její výsledky.
Při pohledu na dendrogram pro metodu průměrné vazby zjistíme, že bude vhodné rozdělit stanice do dvou shluků. Stanice DOB, KRA, HUS a SKA tvoří jeden shluk, zbylých šest stanic druhý shluk. Přitom stanice POL, která se na ploše prvních dvou hlavních komponent poněkud vyčleňovala, se ke 2. shluku skutečně připojí nejpozději.

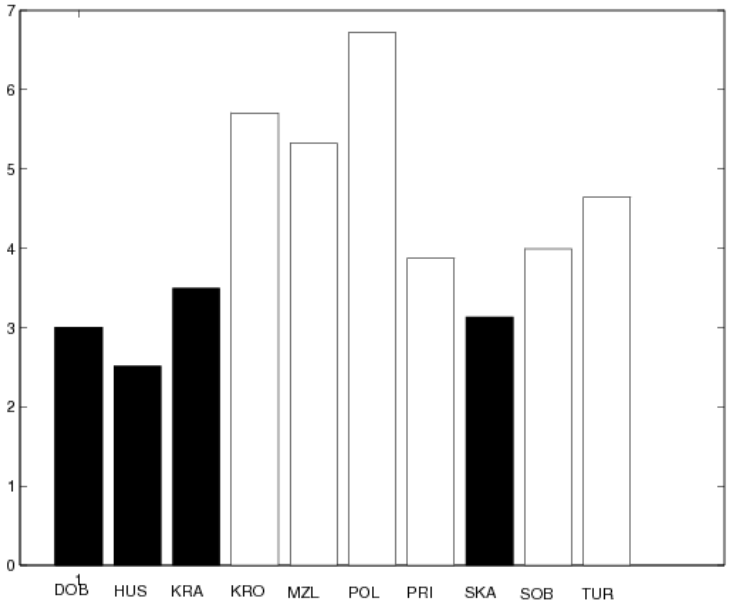
Charakteristiky nalezených shluků

První shluk je tvořen stanicemi, které se vyznačují poměrně nízkými průměrnými ročními koncentracemi oxidu siřičitého (od 6 $\mu g/m^3$ po 11 $\mu g/m^3$ i malými směrodatnými odchylkami (od 2,5 $\mu g/m^3$ po 3,5 $\mu g/m^3$). S výjimkou stanice KRA jsou umístěny v centrální části města.

Druhý shluk obsahuje stanice s vysokými koncentracemi oxidu siřičitého (od 13 $\mu g/m^3$ po 19 $\mu g/m^3$) i poměrně velkými směrodatnými odchylkami (od 3,8 $\mu g/m^3$ po 6,8 $\mu g/m^3$). Tři z nich se nacházejí v okrajových částech Brna (PRI, SOB, TUR), další tři jsou v centru (MZL, KRO, POL).



Sloupkový diagram průměrů



Sloupkový diagram směrodatných odchylek

Výsledek shlukovací procedury, k němuž jsme dospěli, se může jevit poněkud paradoxní. Proč tři stanice (DOB, HUS, SKA) umístěné v centru města vykazují nízké koncentrace SO_2, zatímco jiné tři stanice (MZL, KRO, POL), které se nacházejí rovněž v centru, mají vysoké koncentrace SO_2?

Vysvětlení není jednoznačné. Jak bylo poznamenáno v úvodní části, zkoumané stanice měří koncentrace SO_2 dvěma různými metodami. Přepočet výsledků kolorimetrické metody je do jisté míry subjektivní záležitostí a velmi závisí na zkušenostech laboranta. Na stanicích DOB, HUS, KRA, POL a SKA se používá kolorimetrická metoda, na ostatních gravimetrická.
Vysvětlení může rovněž spočívat v objektivních podmínkách, v nichž se dané stanice nacházejí - např. umístění v krajině, sklon k tvorbě inverzních situací, převládající směr větru apod.

Metoda k-průměrů

Chceme-li verifikovat výsledek dané hierarchické shlukovací metody, můžeme tak učinit např. pomocí metody k-průměrů, což je nehierarchická shlukovací procedura, která vychází z následujícího algoritmu:

Algoritmus:
1. krok: Stanovíme počáteční rozklad množiny n objektů do k shluků. Rozklad zpravidla volíme náhodně.
2. krok: Určíme výběrové centroidy v aktuálních shlucích. (Výběrovým centroidem shluku rozumíme hypotetický objekt, jehož vektor pozorování je roven vektoru výběrových průměrů všech objektů patřících do tohoto shluku.)

3. krok: Pro všechny objekty spočteme jejich vzdálenosti od všech výběrových centroidů. Objekt zařadíme do toho shluku, k jehož výběrovému centroidu má nejblíže. Pokud nedošlo v tomto kroku k žádnému přesunu, považujeme aktuální shluky za definitivní, jinak se vracíme ke 2. kroku.

Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Shlukování metodou k-průměrů – OK – Proměnné r93 až r98 – Shlukovat: Případy (řádky), na záložce Detaily ponecháme implicitní počet shluků 2 – OK. Na záložce Detaily vybereme Členy shluků a vzdálenosti. Dostaneme 2 tabulky, které obsahují názvy stanic v 1. a 2. shluku a vzdálenosti stanic od středu shluku:

	Členy shluku číslo 1 (stanice.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 4 příp.
	Vzdálen.
DOB	0,491653
HUS	0,429539
KRA	0,316674
SKA	0,651282

	Členy shluku číslo 2 (stanice.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 6 příp.
	Vzdálen.
KRO	0,565838
MZL	0,244349
POL	0,828039
PRI	0,376408
SOB	0,381547
TUR	0,807461

Vidíme, že metoda k průměrů dospěla k témuž výsledku jako metoda průměrné vazby.
1. shluk: DOB, KRA, HUS, SKA.
2. shluk: MZL, PRI, SOB, KRO, TUR, POL.
Tento rozklad vyčerpává 67 % variability obsažené v datech.

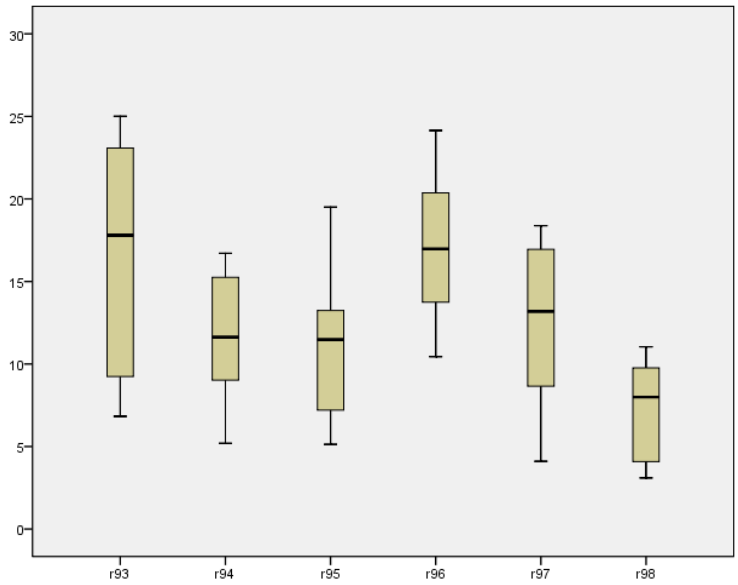
Vliv, který mají jednotlivé proměnné na zařazení do shluků, můžeme posoudit pomocí tabulky ANOVA: na záložce Základní výsledky vybereme Analýza rozptylu:

	Analýza rozptylu (stanice.sta)					
Proměnná	Mezisk. SČ	sv	Vnitřní SČ	sv	F	význam. p
r93	7,180394	1	1,819606	8	31,56900	0,000499
r94	6,069239	1	2,930761	8	16,56700	0,003582
r95	5,691066	1	3,308934	8	13,75928	0,005962
r96	6,453049	1	2,546951	8	20,26910	0,001996
r97	6,567978	1	2,432022	8	21,60500	0,001649
r98	4,305515	1	4,694485	8	7,33714	0,026711

Z hodnoty statistiky F vyplývá, že největší vliv má proměnná r93.

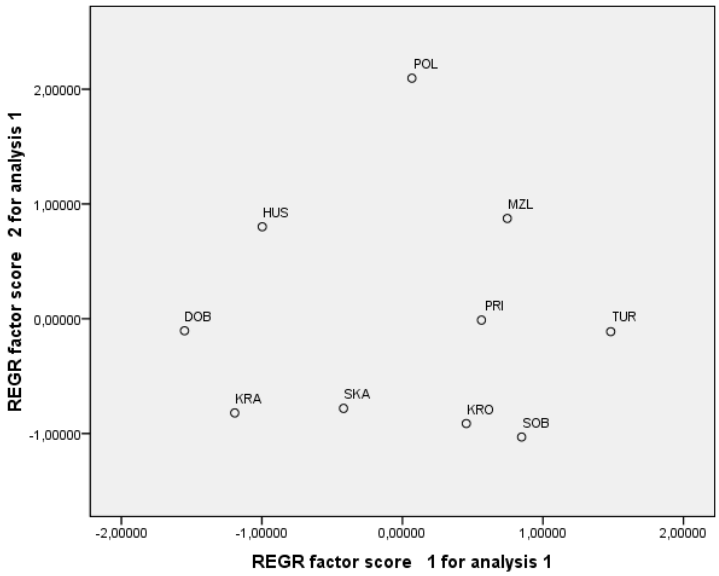
Provedení shlukové analýzy v systému SPSS

Vytvoření krabicových grafů proměnných r93 až r98:
Graphs – Legacy Dialogs – Box plot – ponecháme Simple – zaškrtneme Summaries of separate variables – Define – Boxes Represent r93 až r98 – OK



Vytvoření datového souboru standardizovaných hodnot:
Analyze – Descriptive Statistics – Descriptives – Variables r93 až r98 - zaškrtneme Save stand. Values as variables – OK. Standardizované proměnné se datovém okně objeví jako Zr93 až Zr98.

Vizualizace dat na ploše prvních dvou hlavních komponent:
Analyze – Data Reduction – Factor – Variables – zr93 až zr28 – Etraction – Number of factors – 2 – Scores – Save as variables – Continue.
V datovém souboru se nám objeví dvě nové proměnné FAC1_1 a FAC2_1.
Graphs – Legacy Dialogs – Sccater/Dot – Define – X Axis FAC1_1, Y Axis FAC2_1, Label Cases by stanice – Options – zaškrtneme Display Chart with case labeles – Continue – OK



Provedení shlukové analýzy:

Analyze – Classify – Hierarchical Cluster – Variables Zr93 až Zr98 – Label Cases by stanice – Method Nearest neighbor (metoda nejbližšího souseda), poté Furthest neighbor (metoda nejvzdálenějšího souseda), Between groups linkage (metoda průměrné vazby), Ward's method – Measure Euclidean distance - Continue – Plots – zaškrtneme Dendrogram

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine

```
    C A S E          0         5        10        15        20        25
Label         Num    +---------+---------+---------+---------+---------+

DOB             1
KRA             3
HUS             2
SKA             8
MZL             5
PRI             7
SOB             9
KRO             4
POL             6
TUR            10
```

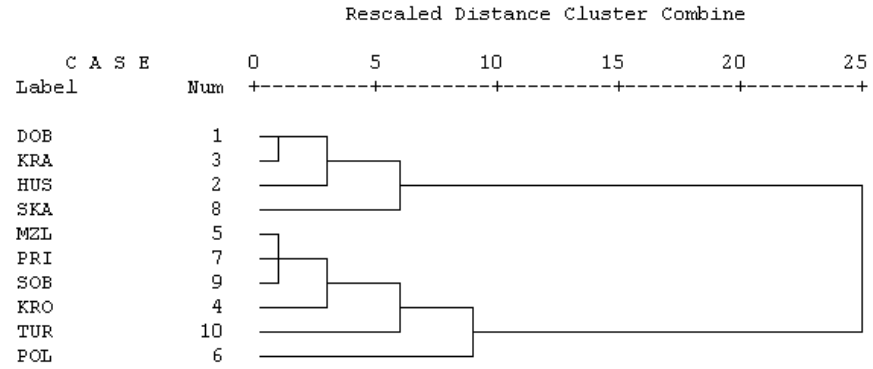
Dendrogram using Complete Linkage

Rescaled Distance Cluster Combine

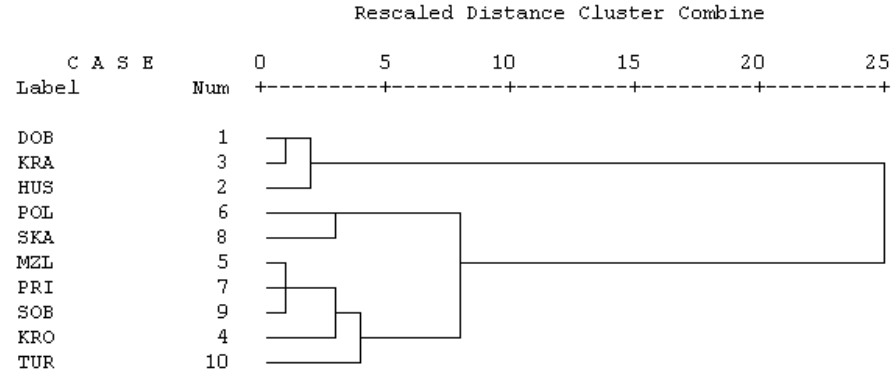
```
    C A S E          0         5        10        15        20        25
Label         Num    +---------+---------+---------+---------+---------+

DOB             1
KRA             3
HUS             2
POL             6
SKA             8
MZL             5
PRI             7
SOB             9
KRO             4
TUR            10
```

Dendrogram using Average Linkage (Between Groups)



Dendrogram using Ward Method



Provedení metody k-průměrů:

Analyze – Classify – K-Means Cluster – Variables Zr93 až Zr98 – Save - zaškrtneme Cluster Membership – Option – zaškrtneme ANOVA table – Continue – OK.

V proměnné QCL_1 se u jednotlivých stanic objeví číslo shluku, do něhož patří.

Tabulka ANOVA:

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(r93)	7,180	1	,227	8	31,569	,000
Zscore(r94)	6,069	1	,366	8	16,567	,004
Zscore(r95)	5,691	1	,414	8	13,759	,006
Zscore(r96)	6,453	1	,318	8	20,269	,002
Zscore(r97)	6,568	1	,304	8	21,605	,002
Zscore(r98)	4,306	1	,587	8	7,337	,027

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The o significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Vliv jednotlivých proměnných na zařazení do shluků můžeme posoudit pomocí statistiky F. Největší vliv má proměnná Zr93.