Cvičení 2.: Shluková analýza

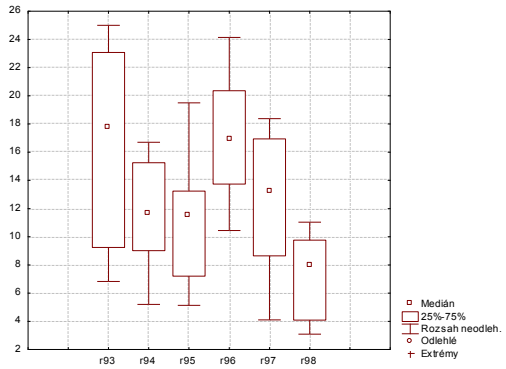
V souboru stanice.sta jsou uloženy údaje (v $\mu g/m^3$) o průměrných ročních koncentracích oxidu siřičitého v letech 1993 – 1998 na deseti brněnských měřicích stanicích: Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice, Tuřany.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

Datový soubor:

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

Úkol 1.: Soubor stanice.sta upravte tak, aby případy 1 až 10 byly pojmenovány názvy stanic.
Návod: Data – Správce jmen případů – Délka jména příp. 5, Přenést jména případů z proměnné Stanice, OK.

Úkol 2.: Prozkoumejte proměnné r93 až r98 pomocí krabicových diagramů.
Návod: Grafy – 2D Grafy – Krabicové grafy –Typ grafu vícenásobný – Proměnné r93, ..., r98, OK, OK.

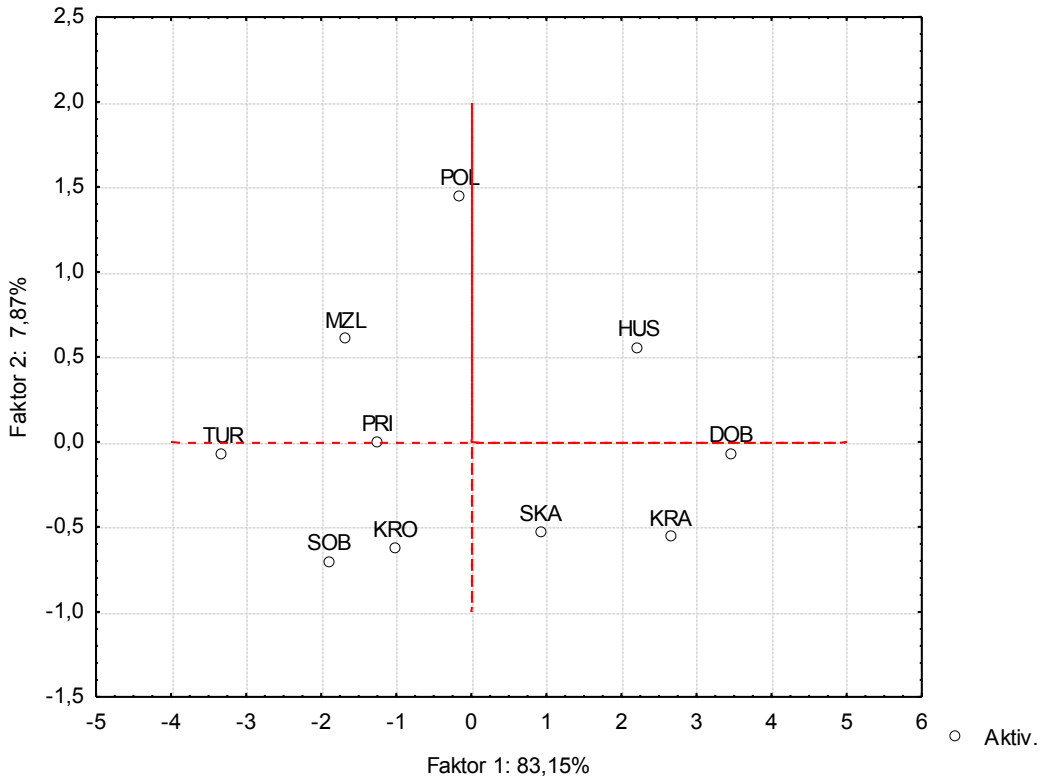


Interpretace: Z krabicových diagramů je vidět, že proměnné r93 až r98 vykazují velmi rozdílnou variabilitu. Nejvyšší variabilitu ve sledovaných deseti stanicích měly koncentrace oxidu siřičitého v roce 1993, naopak nejmenší v roce 1998.

Úkol 3.: Vzhledem k velmi rozdílné variabilitě proměnných r93 až r98 vytvořte standardizované proměnné a nadále pracujte s nimi.
Návod: Data – Standardizovat – Proměnné r93, ..., r98, OK.

	1 Stanice	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
DOB	DOB	-1,398	-1,4569	-1,3398	-1,2048	-1,7224	-1,3635
HUS	HUS	-1,0591	-0,514	-0,1653	-1,4591	-1,1255	-1,11
KRA	KRA	-1,3451	-1,3799	-1,326	-0,714	-0,7964	-1,0339
KRO	KRO	1,01924	-0,5748	0,28819	0,29058	0,61898	0,85957
MZL	MZL	0,88441	1,09043	0,54408	0,78159	0,56013	0,24685
POL	POL	1,15491	0,7081	-0,0589	-0,258	-0,3042	-0,7568
PRI	PRI	-0,1275	0,86598	0,5184	0,57199	0,89228	0,29889
SKA	SKA	-0,349	-0,4657	-0,8647	-0,5575	-0,326	0,29288
SOB	SOB	0,41376	0,5241	0,45007	0,91371	1,01875	1,2855
TUR	TUR	0,80646	1,20277	1,95397	1,63553	1,18432	1,2805

Úkol 4.: Z proměnných r93 až r98 vytvořte dvě hlavní komponenty a graficky znázorněte rozmístění stanic na ploše oprvních dvou hlavních komponent.
Návod: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné r93, ..., r98, OK, OK – Počet faktorů 2, zaškrtneme 2D graf fakt. souřadnic případů.

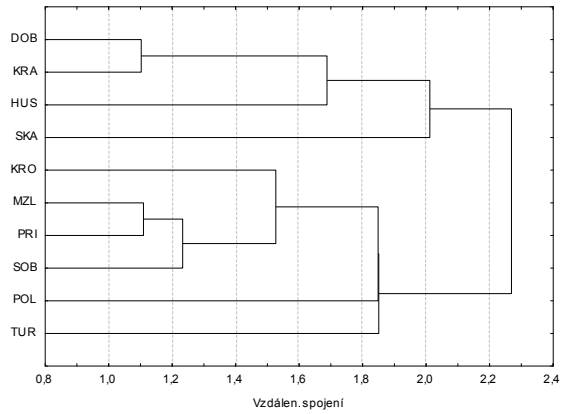


Interpretace: Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

Úkol 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.

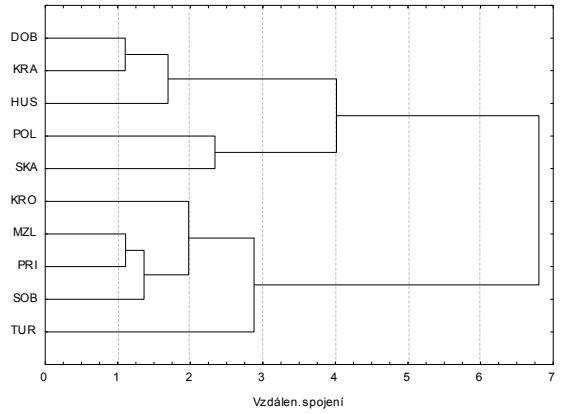
Návod: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné r93 až r98 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednoduché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální graf hierarch. stromu. Pro další dvě metody na záložce Detaily vybereme pravidlo slučování Úplné spojení resp. Nevážený průměr skupin dvojic.

Dendrogram pro metodu nejbližšího souseda



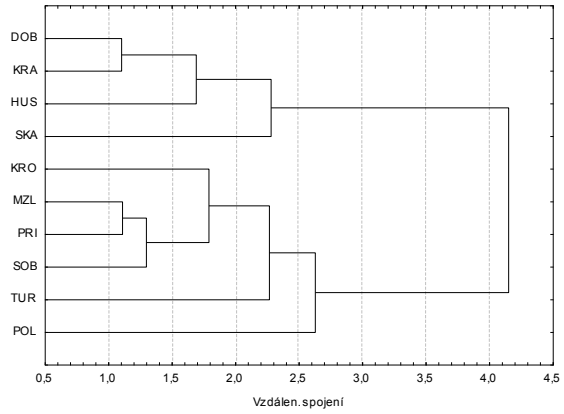
Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, DOB, POL a TUR druhý shluk.

Dendrogram pro metodu nejvzdálenějšího souseda



Interpretace: Stanice DOB, KRA, HUS, POL a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB a TUR druhý shluk.

Dendrogram pro metodu průměrné vazby



Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB, TUR a POL druhý shluk.

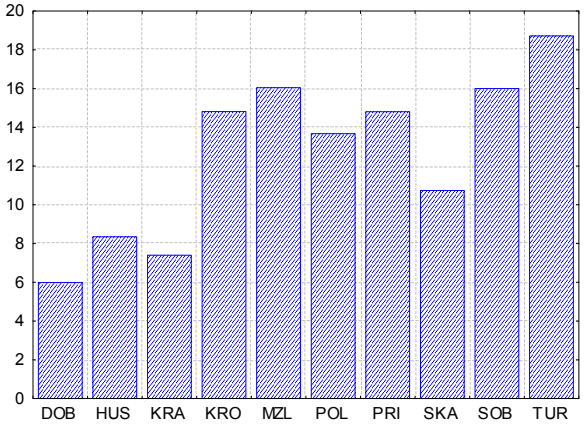
Shrneme-li výsledky všech tří metod, je zřejmé, že stanice DOB, KRA, HUS a STA zřejmě patří do jednoho shluku, zatímco stanice KRO, MZL, SOB a TUR patří do druhého shluku. Příslušnost stanice POL k jednomu či druhému shluku není jednoznačná.

Úkol 6.: Vypočtěte a pomocí sloupkových diagramů znázorněte průměrné roční koncentrace SO_2 a směrodatné odchylky za celé sledované období pro všech deset stanic.
Návod: Je nutné se vrátit k původním nestandardizovaným hodnotám, tj. znovu načíst soubor stanice.sta a pojmenovat případy názvy stanic – viz úkol 1. Pak je zapotřebí soubor transponovat – zaměnit řádky za sloupce: Data – Transponovat – Soubor. Vymažeme 1. řádek: Případy – Odstranit – Od případu 1 do případu 1, OK. Pomocí Popisných statistik vypočteme průměry a směrodatné odchylky proměnných DOB až TUR.

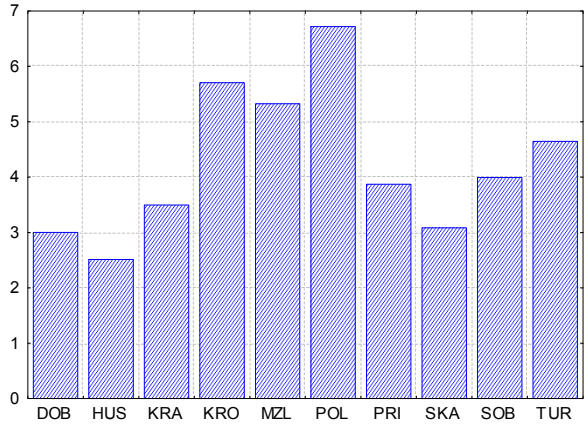
	Popisné statistiky (Tema4
Proměnná	Průměr	Sm. odch.
DOB	5,98933	3,003043
HUS	8,35250	2,513866
KRA	7,40233	3,496625
KRO	14,80767	5,707322
MZL	16,04233	5,326765
POL	13,67333	6,719292
PRI	14,80417	3,873187
SKA	10,74233	3,083617
SOB	15,99900	3,993683
TUR	18,71317	4,645334

Vytvoření sloupkových diagramů pro průměry: ve workbooku klikneme pravým tlačítkem myši na sloupek Průměr: Grafy bloku dat – Vlastní graf bloku podle sloupce – Typ grafu – Sloupcové/pruhové grafy - OK. Podobně pro směrodatné odchylky.

Sloupkový diagram pro průměry



Slopupkový diagram pro sm. 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 $\mu g/m^3$).

Příklad k samostatnému řešení:
U 12 velmi slavných amerických hráčů košíkové byly v sezóně 1989 zjištěny hodnoty osmi proměnných.
Výška – výška hráče v cm
Hmotnost – hmotnost hráče v kg
FgPct – první antropometrická charakteristika
FtPct – druhá antropometrická charakteristika
Body – průměrný počet dosažených bodů
Doskoky - průměrný počet doskoků
Asistence – průměrný počet asistencí
Fauly – průměrný počet faulů
Data jsou uložena v souboru hraci_kosikove.sta.

	1 Jméno hráče	2 Vyska	3 Hmotnost	4 Fgpct	5 Ftpct	6 Body	7 Doskoky	8 Asistence	9 Fauly
1	Jabbar K.A.	218,6	105,0	55,9	72,1	24,6	11,2	3,6	3
2	Barry R.	200,8	93,6	44,9	90,0	23,2	6,7	4,9	3
3	Baylor E.	195,7	102,7	43,1	78,0	27,4	13,5	4,3	3,1
4	Bird L.	205,9	100,4	50,3	88,0	25,0	10,2	6,1	2,7
5	Chamberlain W.	216,0	125,5	54,0	51,1	30,1	22,9	4,4	2
6	Cousy B.	184,3	79,9	37,5	80,3	18,4	5,2	7,5	2,4
7	Erving J.	199,5	91,3	50,6	77,8	24,2	8,5	4,2	2,8
8	Johnson M.	205,9	98,1	53,0	83,4	19,5	7,4	11,2	2,4
9	Jordan M.	198,3	89,0	51,3	84,8	32,6	6,2	5,9	3,1
10	Robertson O.	195,7	95,8	48,5	83,8	25,7	7,5	9,5	2,8
11	Russell B.	207,1	100,4	44,0	56,1	15,1	22,6	4,3	2,7
12	West J.	189,4	82,2	47,4	81,4	27,0	5,8	6,7	2,6

Metodami shlukové analýzy najděte skupiny hráčů podobných vlastností.

(Příklad je převzat z knihy M. Meloun, J. Militký, M. Hill: Počítačová analýza vícerozměrných dat. Academia Praha 2005)