Cvičení 2.: Shluková analýza

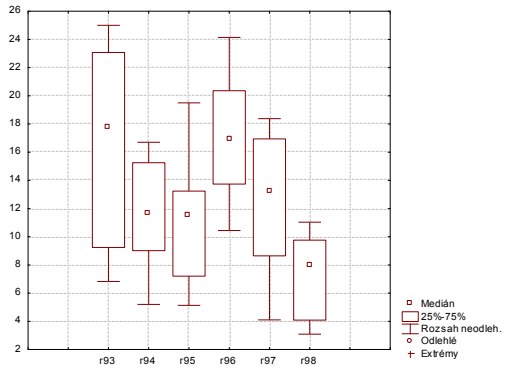
V souboru stanice.sta jsou uloženy údaje (v μg/m^3) o průměrných ročních koncentracích oxidu siřičitého v letech 1993 – 1998 na deseti brněnských měřicích stanicích: Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice, Tuřany. Cílem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

Datový soubor:

	1 Stani	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
1	DOB	6,8	5,2	5,1	11,5	4,1	3,0
2	HUS	9,2	9,2	10,2	10,4	7,0	3,8
3	KRA	7,2	5,5	5,1	13,7	8,6	4,0
4	KRO	24,0	9,0	12,2	18,1	15,6	9,7
5	MZL	23,0	16,2	13,3	20,3	15,3	7,9
6	POL	25,0	14,5	10,7	15,	11,0	4,9
7	PRI	15,8	15,2	13,2	19,4	16,9	8,0
8	SKA	14,2	9,4	7,2	14,4	10,9	8,0
9	SOB	19,7	13,7	12,9	20,9	17,5	11,0
10	TUR	22,5	16,7	19,5	24,1	18,3	11,0

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.

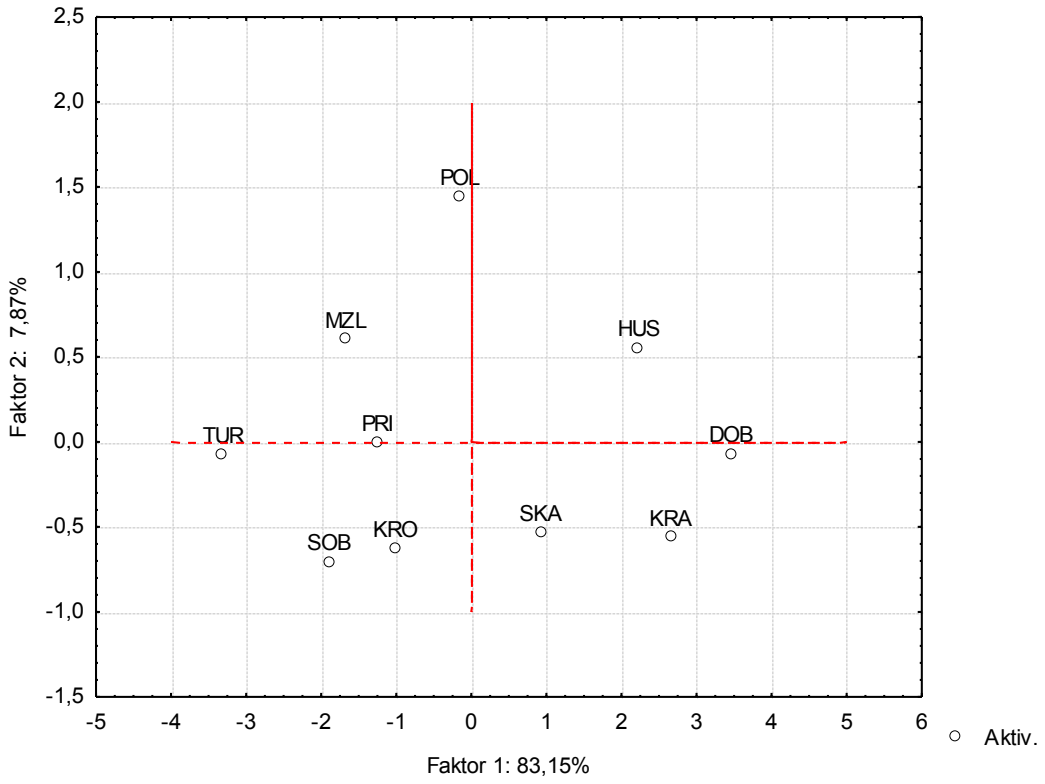


Interpretace: Z krabicových diagramů je vidět, že proměnné r93 až r98 vykazují velmi rozdílnou variabilitu. Nejvyšší variabilitu ve sledovaných deseti stanicích měly koncentrace oxidu siřičitého v roce 1993, naopak nejmenší v roce 1998.

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 Stani	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
DOB	DOB	-1,3	-1,45	-1,33	-1,20	-1,72	-1,36
HUS	HUS	-1,05	-0,5	-0,16	-1,45	-1,12	-1,
KRA	KRA	-1,34	-1,37	-1,3	-0,7	-0,79	-1,03
KRO	KRO	1,019	-0,57	0,288	0,290	0,618	0,859
MZL	MZL	0,884	1,090	0,544	0,781	0,560	0,240
POL	POL	1,154	0,70	-0,05	-0,2	-0,30	-0,75
PRI	PRI	-0,12	0,865	0,51	0,57	0,892	0,298
SKA	SKA	-0,3	-0,46	-0,86	-0,55	-0,3	0,292
SOB	SOB	0,413	0,52	0,450	0,913	1,018	1,28
TUR	TUR	0,800	1,202	1,953	1,635	1,184	1,28

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.

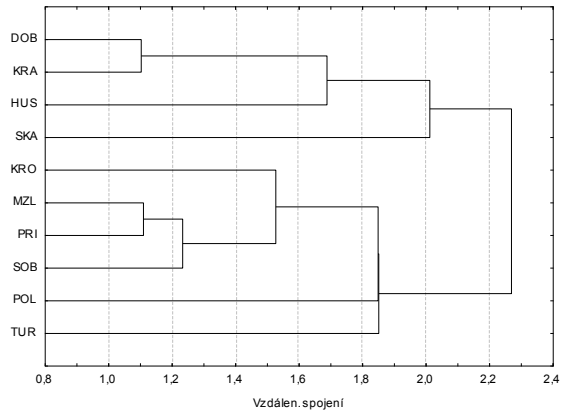


Interpretace: Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

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.

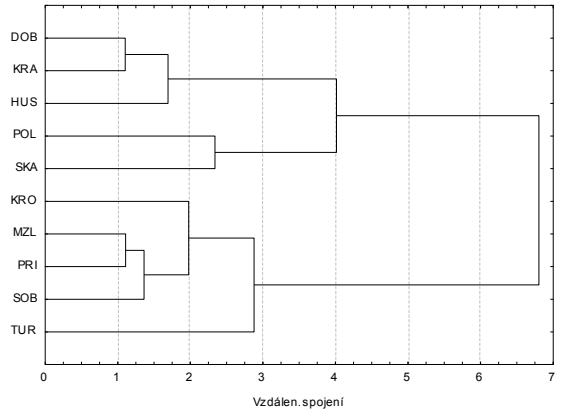
Návod: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné r93 až r98 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednoduché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální 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



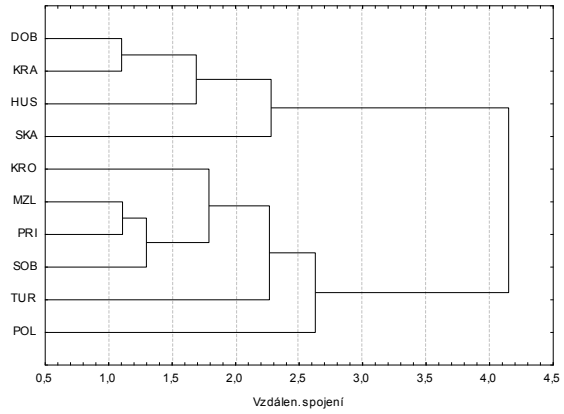
Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, DOB, POL a TUR druhý shluk.

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.

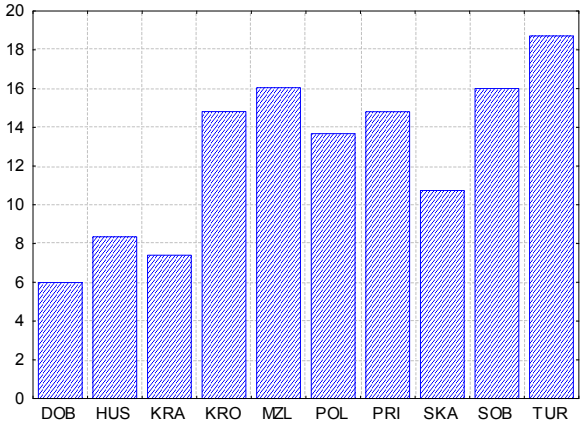
Shrneme-li výsledky všech tří metod, je zřejmé, že stanice DOB, KRA, HUS a STA zřejmě patří do jednoho shluku, zatímco stanice KRO, MZL, SOB a TUR patří do druhého shluku. Příslušnost stanice POL k jednomu či druhému shluku není jednoznačná.

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.

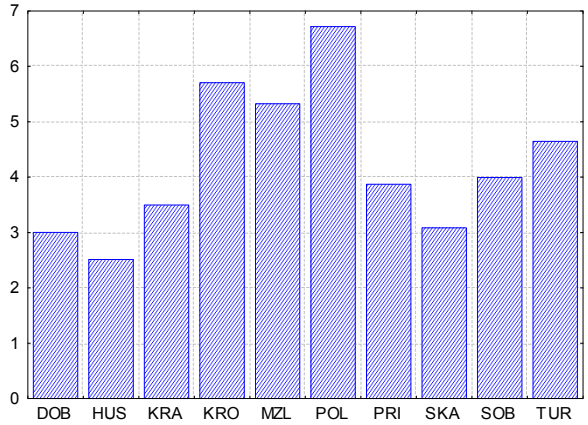
	Popisné statistiky (Tema4)	
Proměnná	Průměr	Sm. odch.
DOB	5,98933	3,003043
HUS	8,35250	2,513866
KRA	7,40233	3,496625
KRO	14,80767	5,707322
MZL	16,04233	5,326765
POL	13,67333	6,719292
PRI	14,80417	3,873187
SKA	10,74233	3,083617
SOB	15,99900	3,993683
TUR	18,71317	4,645334

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 hr	2 Výsk	3 Hmotn	4 Fgp	5 Ftp	6 Boc	7 Dosk	8 Asister	9 Fau
1	Jabbar K.A	218	105	55	72	24	11	3,	3
2	Barry R.	200	93	44	90	23	6,	4,	3
3	Baylor E.	195	102	43	78	27	13	4,	3,
4	Bird L.	205	100	50	88	25	10	6,	2,
5	Chamberla	216	125	54	51	30	22	4,	2
6	Cousy B.	184	79	37	80	18	5,	7,	2,
7	Erving J.	199	91	50	77	24	8,	4,	2,
8	Johnson M	205	98	53	83	19	7,	11	2,
9	Jordan M.	198	89	51	84	32	6,	5,	3,
10	Robertson	195	95	48	83	25	7,	9,	2,
11	Russell B.	207	100	44	56	15	22	4,	2,
12	West J.	189	82	47	81	27	5,	6,	2,

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)