Shluková analýza

Motivace:

S problematikou klasifikace objektů do skupin se v praxi setkáváme velmi často. Např. biolog studuje 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.“

Shluky můžeme popsat jako "nepřerušované oblasti prostoru obsahující relativně velkou hustotu bodů, oddělených od dalších takových oblastí oblastmi, které obsahují relativně malou hustotu bodů. Důležitost tohoto popisu je v tom, že předtím než se uskuteční analýza dat, neomezuje chápání tohoto pojmu na žádnou konkrétní podobu.

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.

Příklad:
Uvažme datový soubor, který vznikl tak, že 6 žáků absolvovalo 4 testy, které měří následující veličiny:
X_1 – přírodovědné znalosti,
X_2 – literární vědomosti,
X_3 – schopnost koncentrace,
X_4 – logické myšlení.
Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek)

	1 X1	2 X2	3 X3	4 X4
1	7	9	1	8
2	9	8	8	1
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

Vypočtěte matici euklidovských vzdáleností.

Řešení:
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1 – X4 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky) – OK – na záložce Detaily vybereme Matice vzdáleností.

	Euklid. vzdálenosti					
Příp	P	P	P	P	P	P
P_1	0,	3,	12	12	12	14
P_2	3,	0,	12	13	12	14
P_3	12	12	0,	2,	3,	4,
P_4	12	13	2,	0,	3,	3,
P_5	12	12	3,	3,	0,	2,
P_6	14	14	4,	3,	2,	0,

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 čtyři z nich.

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

Nevýhoda: řetězový efekt (spojují se shluky, jejichž dva objekty jsou sice nejbližší, ale vzhledem k většině ostatních objektů nejde o nejbližší shluky)

Výhody: Je invariantní k monotónním transformacím matice podobností a není ovlivněna vazbami v datech. První vlastnost, invariantnost k monotónní transformaci, je celkem důležitá, neboť téměř všechny další hierarchické aglomerativní metody tuto vlastnost nemají. To znamená, že metoda nejbližšího souseda je jedna z mála metod, které nejsou ovlivněny žádnou transformací dat.

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

Výhoda: odpadá řetězový efekt, vede k tvorbě relativně malého počtu poměrně kompaktních shluků.

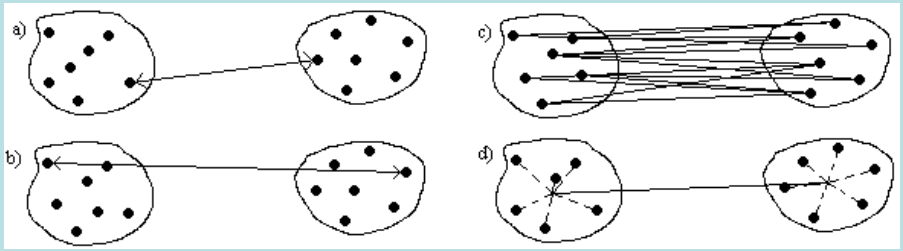
c) Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.

Vede k podobným výsledkům jako metoda nejvzdálenějšího souseda.

Tyto tři metody nevyžadují původní data, stačí jim matice vzdáleností.

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.

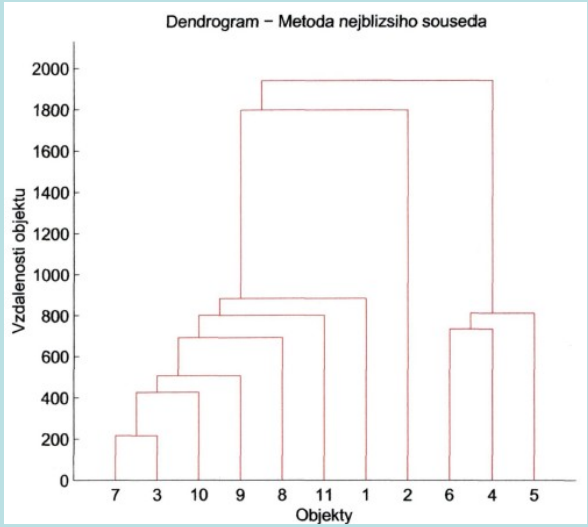
Požaduje vyjádření vzdálenosti objektů čtvercovou euklidovskou vzdáleností.



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.

Na svislé ose připravíme stupnici pro hladiny spojování. Dole začíná strom n větvemi a v každém kroku spojíme dvě větve v bodě, který odpovídá příslušné hladině spojení.



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.

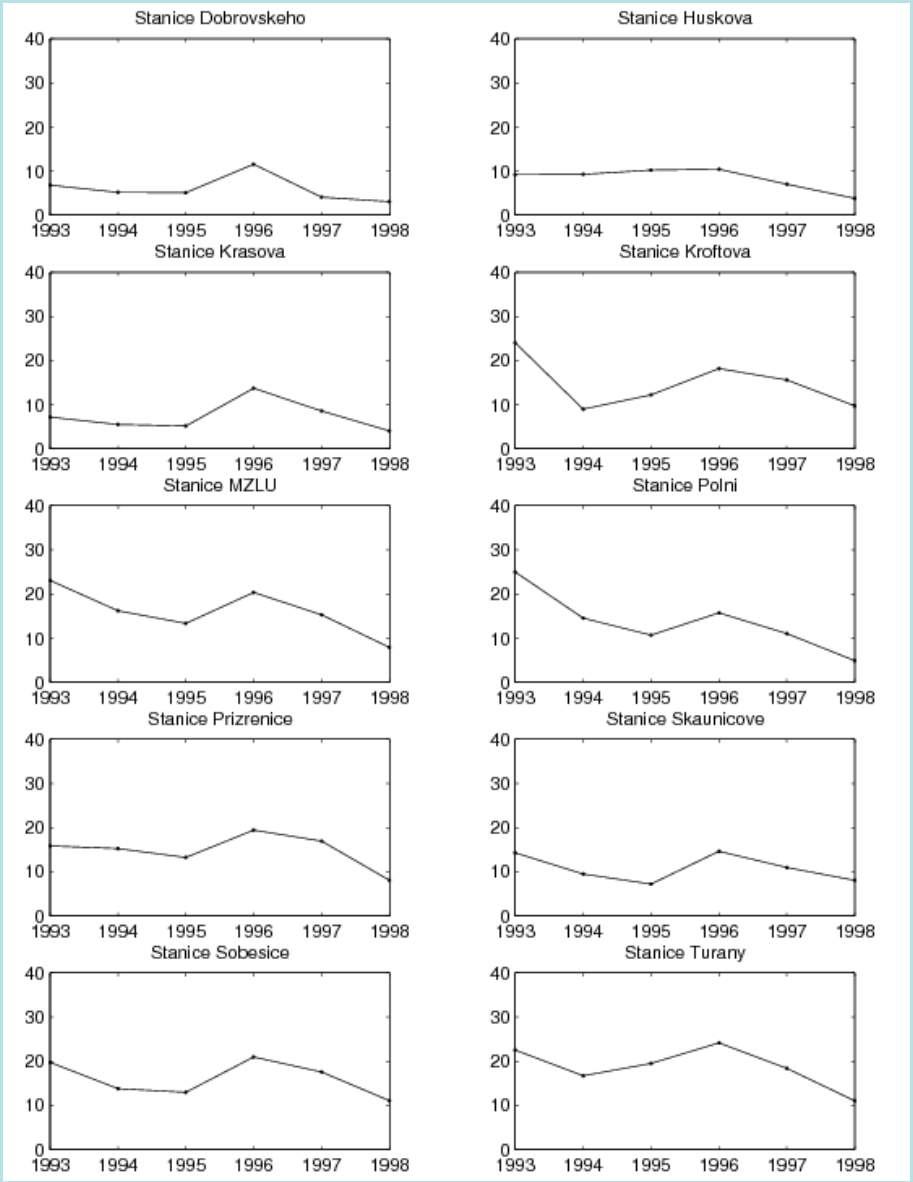
Příklad: Tento příklad vychází z publikace
Budíková, Marie. Aplikace shlukové analýzy v ekologii. Praha : Jednota českých matematiků a fyziků, 2001. 8 s. Sborník prací 11. letní školy ROBUST 2000.
V rámci jedné z bakalářských prací obhájených na katedře geografie byly shromážděny údaje o průměrných měsíčních 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.

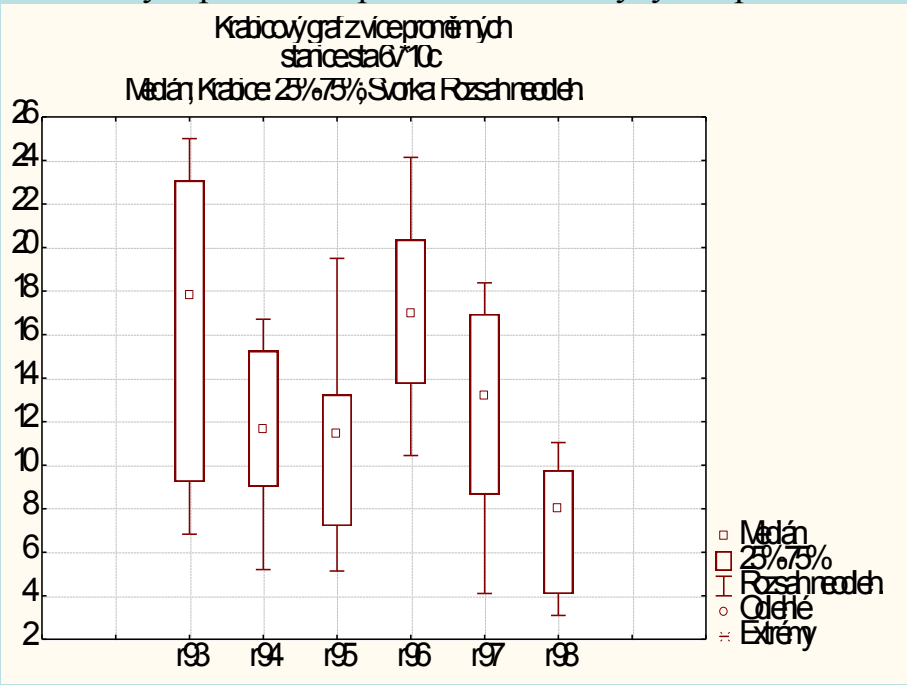
Každá ze sledovaných 10 stanic byly popsána šesti údaji, jak vidíme v této tabulce.

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

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


Stanice Dobrovskeho
Stanice Huskova
Stanice Krasova
Stanice Kroftova
Stanice MZLU
Stanice Polni
Stanice Prizrenice
Stanice Skaunicove
Stanice Sobesice
Stanice Turany
0
10
20
30
40
1993
1994
1995
1996
1997
1998

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

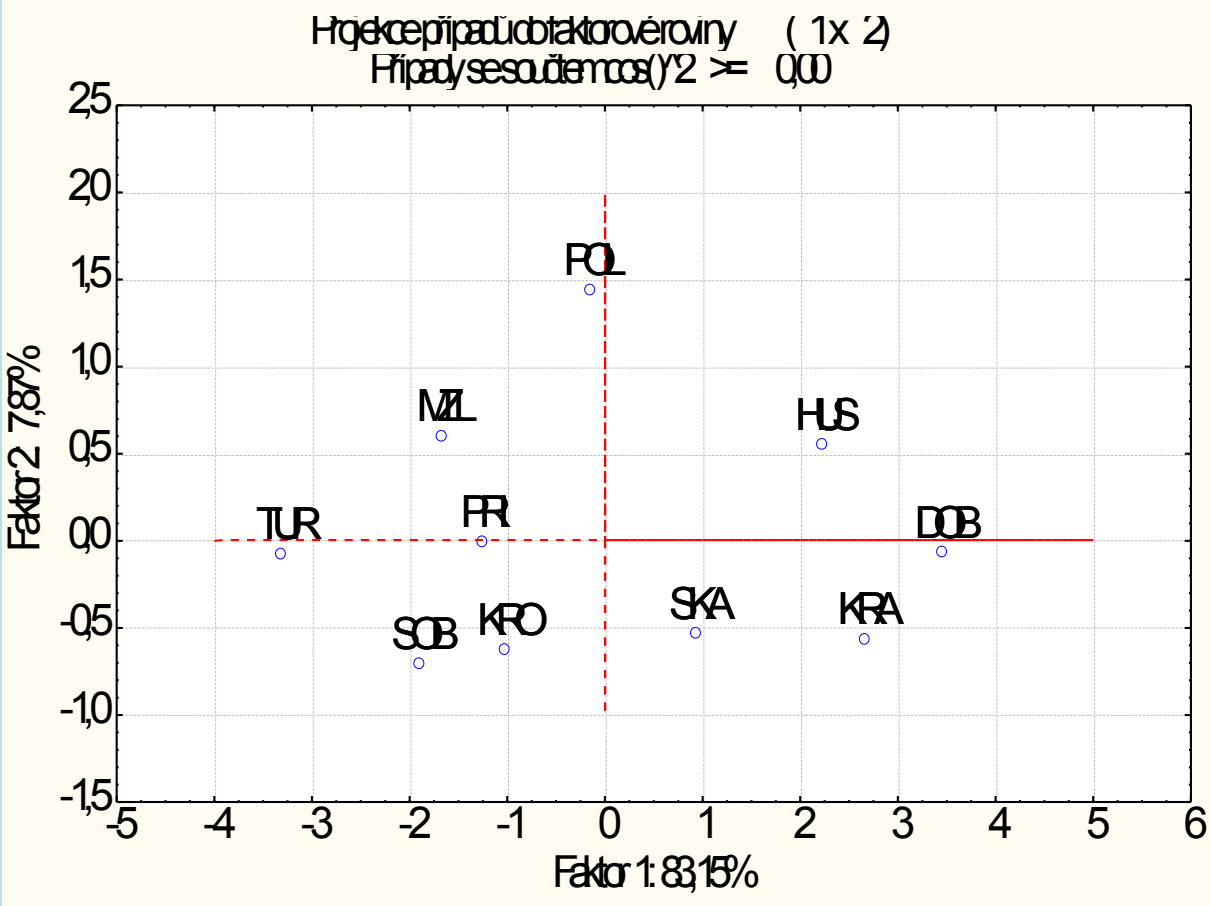
	1 r93	2 r94	3 r95	4 r96	5 r97	6 r98
DOB	-1,3	-1,4	-1,3	-1,2	-1,7	-1,3
HUS	-1,0	-0,5	-0,1	-1,4	-1,1	-1,1
KRA	-1,3	-1,3	-1,3	-0,7	-0,7	-1,0
KRO	1,01	-0,5	0,28	0,29	0,6	0,85
MZL	0,88	1,09	0,54	0,78	0,56	0,24
POL	1,15	0,70	-0,0	-0,2	-0,3	-0,7
PRI	-0,1	0,8	0,51	0,5	0,89	0,29
SKA	-0,3	-0,4	-0,8	-0,5	-0,3	0,29
SOB	0,41	0,52	0,45	0,91	1,01	1,28
TUR	0,80	1,20	1,9	1,63	1,18	1,28

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

	Vlastní čísla korelační matice a souvisejí Pouze aktiv. proměnné			
Pořadí	vl. čís	% ce rozpty	Kumula vl. čísl	Kumula %
1	4,989	83,15	4,989	83,15
2	0,472	7,87	5,461	91,02
3	0,300	5,014	5,762	96,04
4	0,129	2,165	5,892	98,20
5	0,073	1,219	5,965	99,42
6	0,034	0,574	6,000	100,0

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

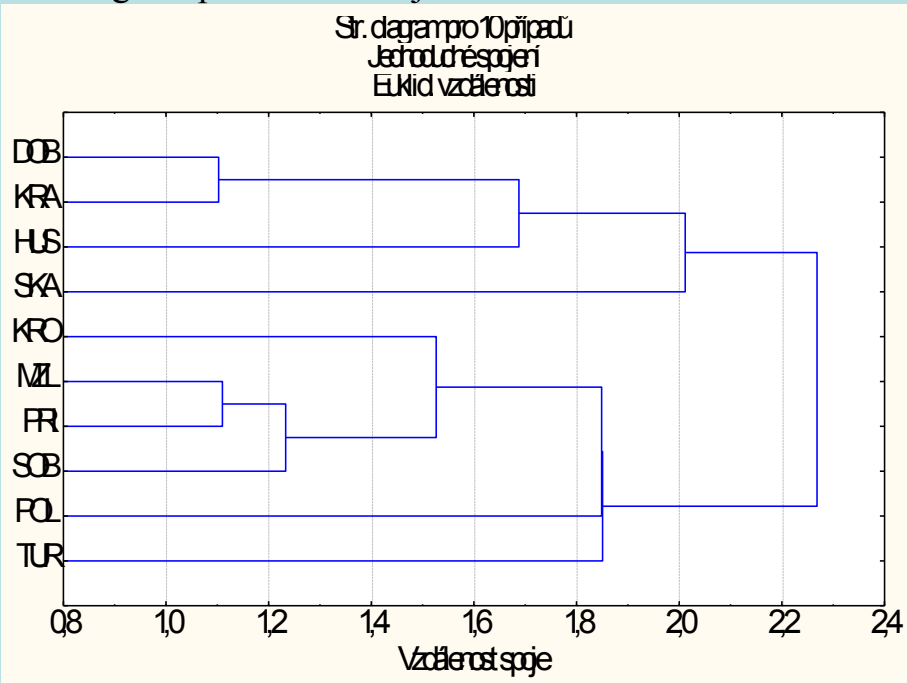
Rozmístění stanic na ploše prvních dvou hlavních komponent:



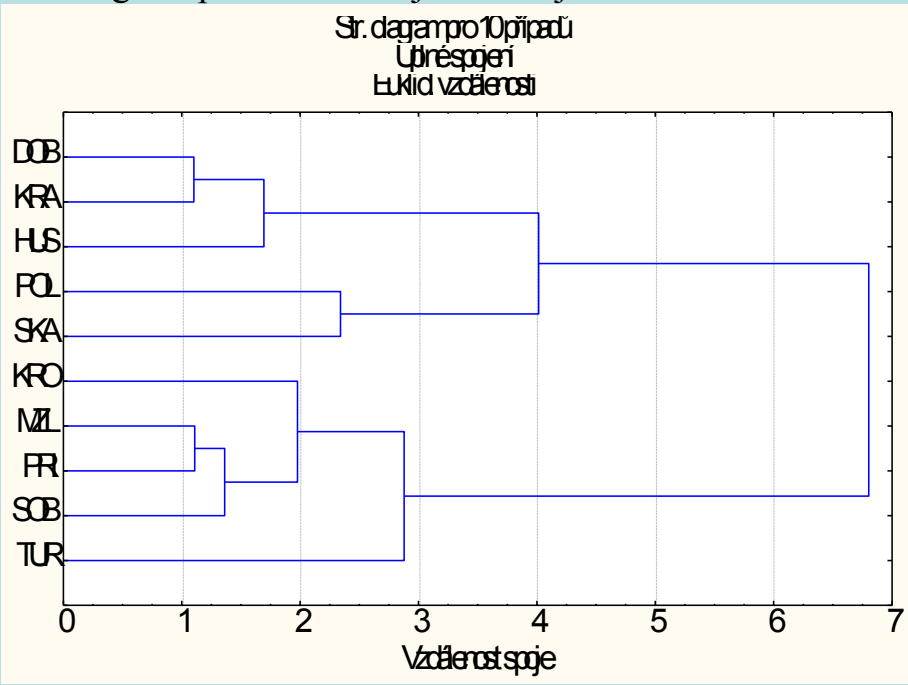
Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

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.
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 změníme Pravidlo slučování z Jednoduchého spojení na Úplné spojení resp. Nevážený průměr skupin dvojic resp. Wardova metoda.

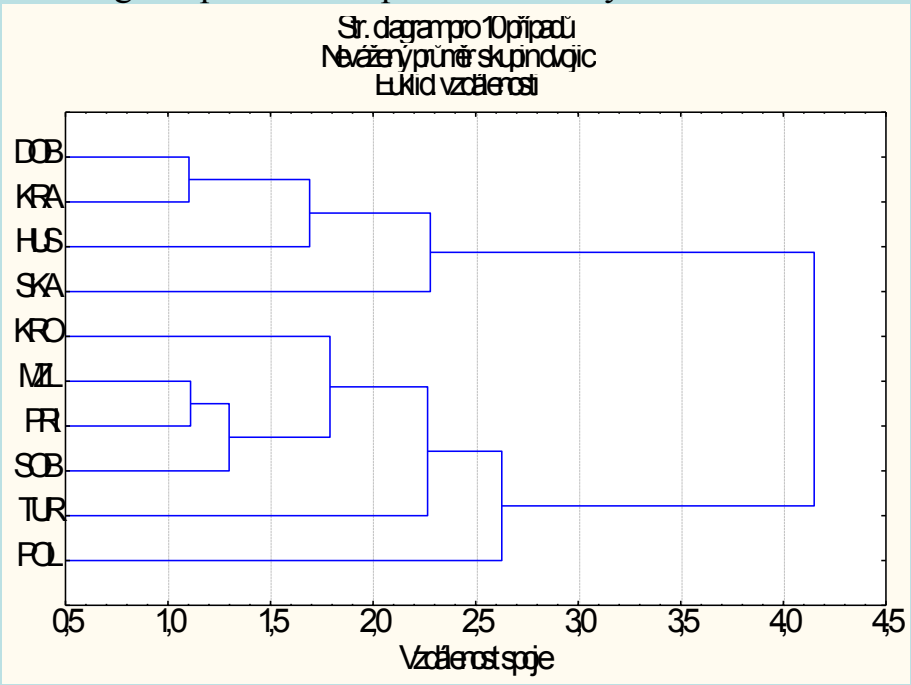
Dendrogram pro metodu nejbližšího souseda:



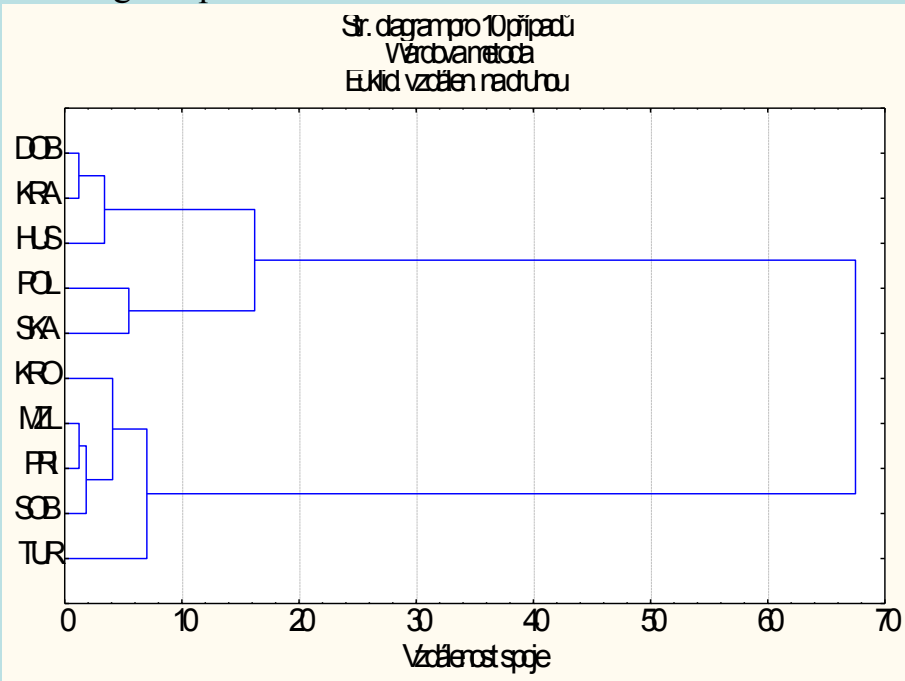
Dendrogram pro metodu nejvzdálenějšího souseda:



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.
Průběh shlukování vidíme na tzv. rozvrhu shlukování:

Amalgamation Schedule (stanice.sta)
Unweighted pair-group average
Euclidean distances

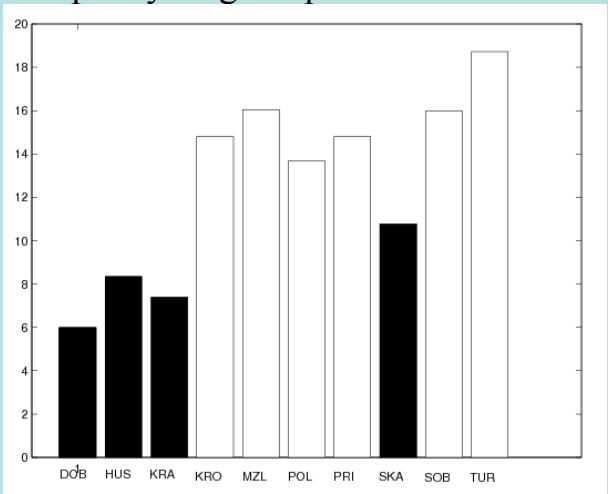
linkag distan	Obj. N 1	Obj. N 2	Obj. N 3	Obj. N 4	Obj. N 5	Obj. N 6	Obj. N 7	Obj. N 8	Obj. N 9	Obj. N 10
1,102	DC	KR								
1,109	MA	PH								
1,298	MA	PH	SC							
1,690	DC	KR	HU							
1,789	KR	MA	PH	SC						
2,265	KR	MA	PH	SC	TU					
2,279	DC	KR	HU	SK						
2,627	KR	MA	PH	SC	TU	PC				
4,150	DC	KR	HU	SK	KR	MA	PH	SC	TU	PC

Charakteristiky nalezených shluků

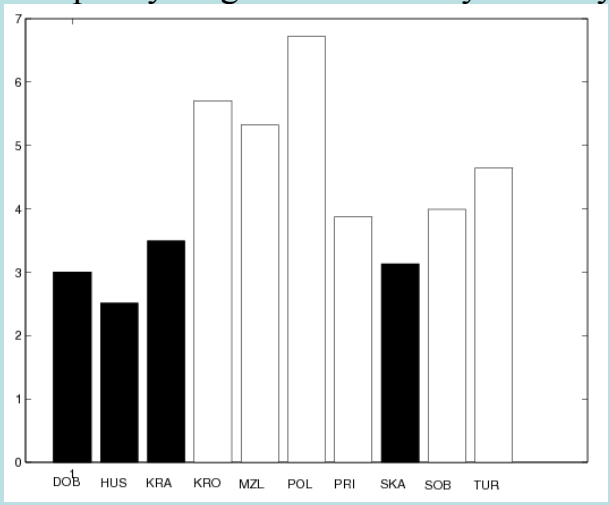
První shluk je tvořen stanicemi, které se vyznačují poměrně nízkými průměrnými ročními koncentracemi oxidu siřičitého (od 6 μg/m^3 po 11 μg/m^3 i malými směrodatnými odchylkami (od 2,5 μg/m^3 po 3,5 μg/m^3). S výjimkou stanice KRA jsou umístěny v centrální části města.

Druhý shluk obsahuje stanice s vysokými koncentracemi oxidu siřičitého (od 13 μg/m^3 po 19 μg/m^3) i poměrně velkými směrodatnými odchylkami (od 3,8 μg/m^3 po 6,8 μg/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.

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 (stani a vzdálenosti od příslušné Shluk obsahuje 4 příp.
	Vzdále
DOB	0,491
HUS	0,429
KRA	0,316
SKA	0,651

	Členy shluku číslo 2 (stani a vzdálenosti od příslušné Shluk obsahuje 6 příp.
	Vzdále
KRO	0,565
MZL	0,244
POL	0,828
PRI	0,376
SOB	0,381
TUR	0,807

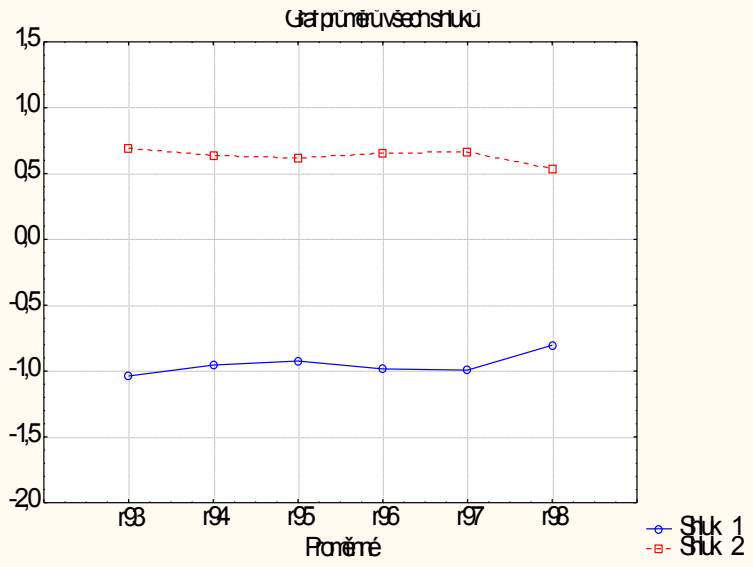
Vidíme, že metoda k průměrů dospěla k témuž výsledku jako metoda průměrné vazby.
1. shluk: DOB, KRA, HUS, SKA.
2. shluk: MZL, PRI, SOB, KRO, TUR, POL.
Tento rozklad vyčerpává 67 % variability obsažené v datech.

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

Proměnná	Analýza rozptylu (stanice.sta)					
	Mezis SC	sv	Vnitřní SC	sv	F	významná p
r93	7,180	1	1,819	8	31,56	0,000
r94	6,069	1	2,930	8	16,56	0,003
r95	5,691	1	3,308	8	13,75	0,005
r96	6,453	1	2,546	8	20,26	0,001
r97	6,567	1	2,432	8	21,60	0,001
r98	4,305	1	4,694	8	7,337	0,026

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

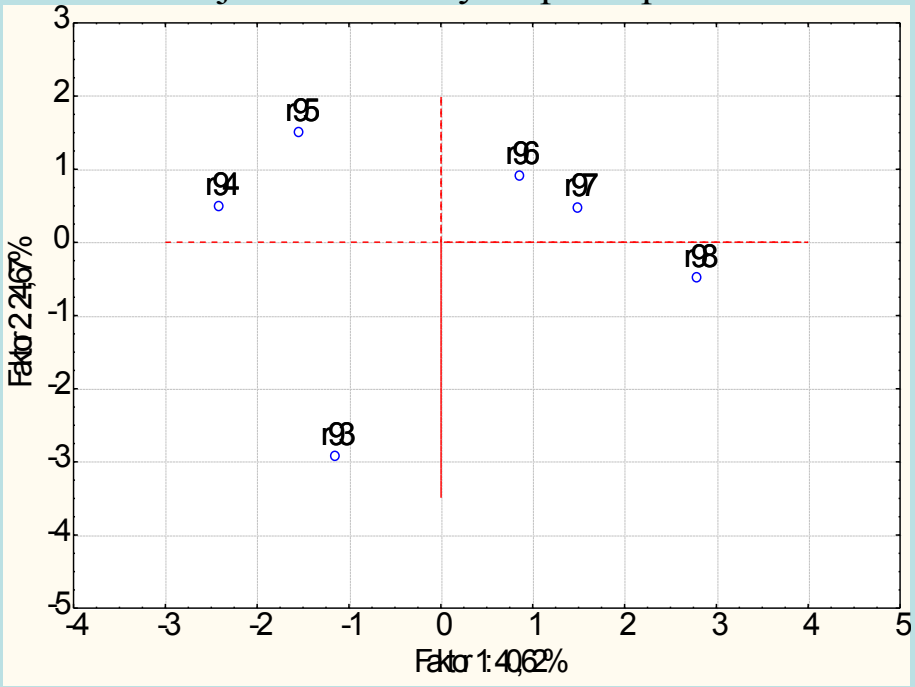
Graf průměrů obou shluků



Shlukování proměnných

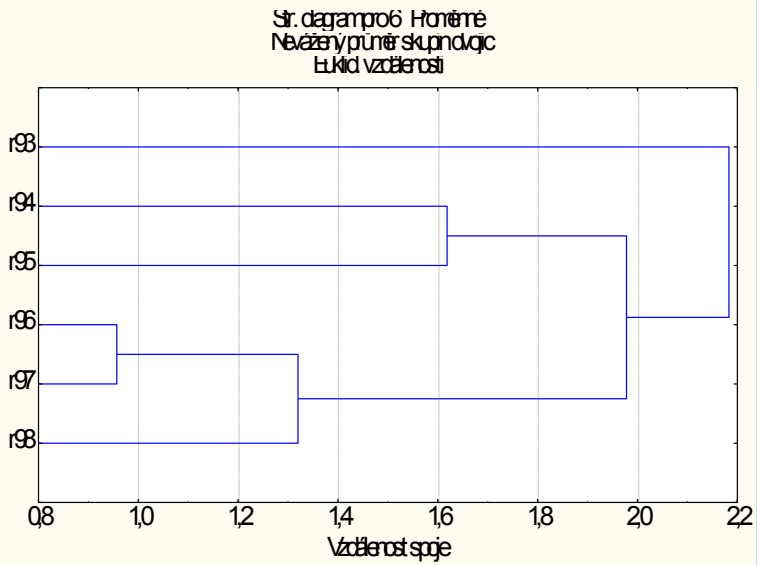
Systém STATISTICA pomocí shlukové analýzy umožňuje zjistit, které proměnné mají k sobě blízko. Budeme pracovat se standardizovanými hodnotami datového souboru stanice.sta.

Znázorníme jednotlivé roky na ploše prvních dvou hlavních komponent:



Je vidět, že blízko k sobě mají proměnné r94, r95, dále r96, r97, r98 a proměnná r93 zůstává stranou.

Provedeme shlukovou analýzu s euklidovskými vzdálenostmi a metodou průměrné vazby.
Výsledky znázorníme dendrogramem.



Provedeme-li řez na úrovni spojení 1,8, dostaneme tři shluky: (r93), (r94, r95) a (r96, r97, r98).
Tento výsledek ještě ověříme metodou k-průměrů pro k = 3.

| | Členy shluku číslo 1 (stan
a vzdálenosti od příslušné
Shluk obsahuje 3 příp. |
|---|---|
| | Vzdále |
| r96 | 0,231 |
| r97 | 0,162 |
| r98 | 0,261 |

| | Členy shluku číslo 2 (stan
a vzdálenosti od příslušné
Shluk obsahuje 2 příp. |
|---|---|
| | Vzdále |
| r94 | 0,255 |
| r95 | 0,255 |

| | Členy shluku číslo 3 (stan
a vzdálenosti od příslušné
Shluk obsahuje 1 příp. |
|---|---|
| | Vzdále |
| r93 | 0,(|

Výsledek metody k-průměrů je v souladu s výsledkem metody průměrné vazby.

Vliv jednotlivých stanic na zařazení roků do shluků posoudíme pomocí tabulky ANOVA:

	Analýza rozptylu (stanice.sta)					
Proměn	Mezis SČ	sv	Vnitřn SČ	sv	F	významná p
DOB	0,001	2	0,147	3	0,015	0,984
HUS	0,982	2	0,138	3	10,63	0,043
KRA	0,378	2	0,056	3	10,01	0,046
SKA	0,264	2	0,466	3	0,850	0,509
KRO	1,080	2	0,535	3	3,025	0,190
MZL	0,147	2	0,293	3	0,752	0,543
POL	2,085	2	0,446	3	7,017	0,073
PRI	0,490	2	0,236	3	3,107	0,185
SOB	0,561	2	0,076	3	11,05	0,041
TUR	0,400	2	0,395	3	1,519	0,350

Na hladině významnosti 0,05 se pro zařazení roků do shluků jeví jako významné stanice HUS, KRA, SOB.

Analýza turistického ruchu ve 23 státech EU

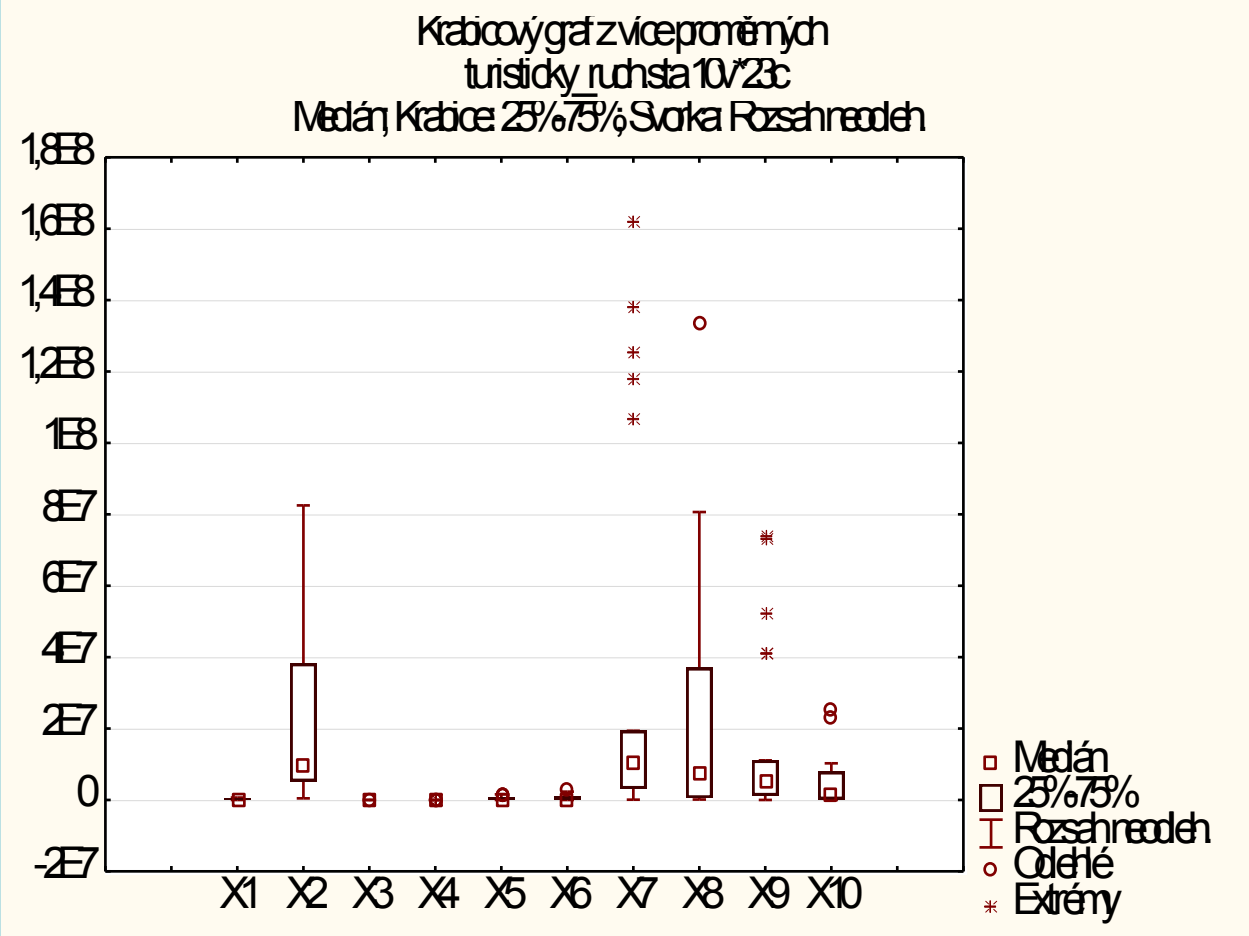
Máme k dispozici datový soubor z EUROSTATu, který popisuje některé vybrané ukazatele turistického ruchu v r. 2005:

1 Stat	2 X1	3 X2	4 X3	5 X4	6 X5	7 X6	8 X7	9 X8	10 X9	11 X10
Belgie	305	10445	189	153	1210	2950	43130	85140	23640	22080
Bulharsko	1109	7761(	123	32	2010	200	39570	4900	17210	1730
Česká rep	788	10220	427	332	2320	2010	86010	12124	33880	26370
Dánsko	430	5411	48	60	700	3230	53160	11556	18990	16240
Estonsko	452	1347	31	46	250	130	7510	3780	4280	1910
Finsko	3381	5236(	93	45	1180	930	10388	2372(	5948(	10610
Francie	6748	62637	198	924	1740(	3039(	125216	62426	73066	10291
Itálie	3013	58462	335	964	2028(	2322(	138222	68504	41295	89180
Litva	652	3425	33	19	200	110	7280	4940	3470	1580
Lotyšsko	645	2306	33	8	190	500	7960	2250	3540	710
Lucembursk	2586	4612	29	25	140	520	850	1450	290	340
Maďarsko	930	10097	206	105	1620	1670	66220	23360	27780	8390
Německo	3570	82500	3650	187	16210	16960	161895	133840	73777	25296
Nizozemí	415	16305	313	402	1920	9980	14375	40575	83010	78810
Polsko	3126	38173	220	452	1700	4000	12464	25612	68050	54820
Portugalsk	923	10529	207	28	2640	1830	11648	62300	52740	12140
Rakousko	838	8206(	142	628	5710	3550	19383	79150	68960	15320
Řecko	1319	11082	903	34	6820	960	13942	5870	59330	1310
Slovensko	490	5384	88	113	570	1030	31830	26380	12440	6560
Slovinsko	202	1997	34	35	300	350	16530	14050	4590	3530
Španělsko	5040	43038	1760	171	15800	14840	106875	36999	41600	85520
Švédsko	4499	9011	185	208	1970	5370	17518	17345	11096	65860
Velká Brit	2448	60059	3290	338	10620	11630	117920	80635	52611	23069

X1 … rozloha, X2 … počet obyvatel, X3 … počet hotelů, X4 … počet jiných ubytovacích zařízení, X5 resp. X6 … počet postelí v hotelech resp. jiných ubytovacích zařízeních, X7 resp. X8 … počet nocí strávených v hotelech resp. jiných ubytovacích zařízeních, X9 resp. X10 … počet příchodů do hotelů resp. jiných ubytovacích zařízení.

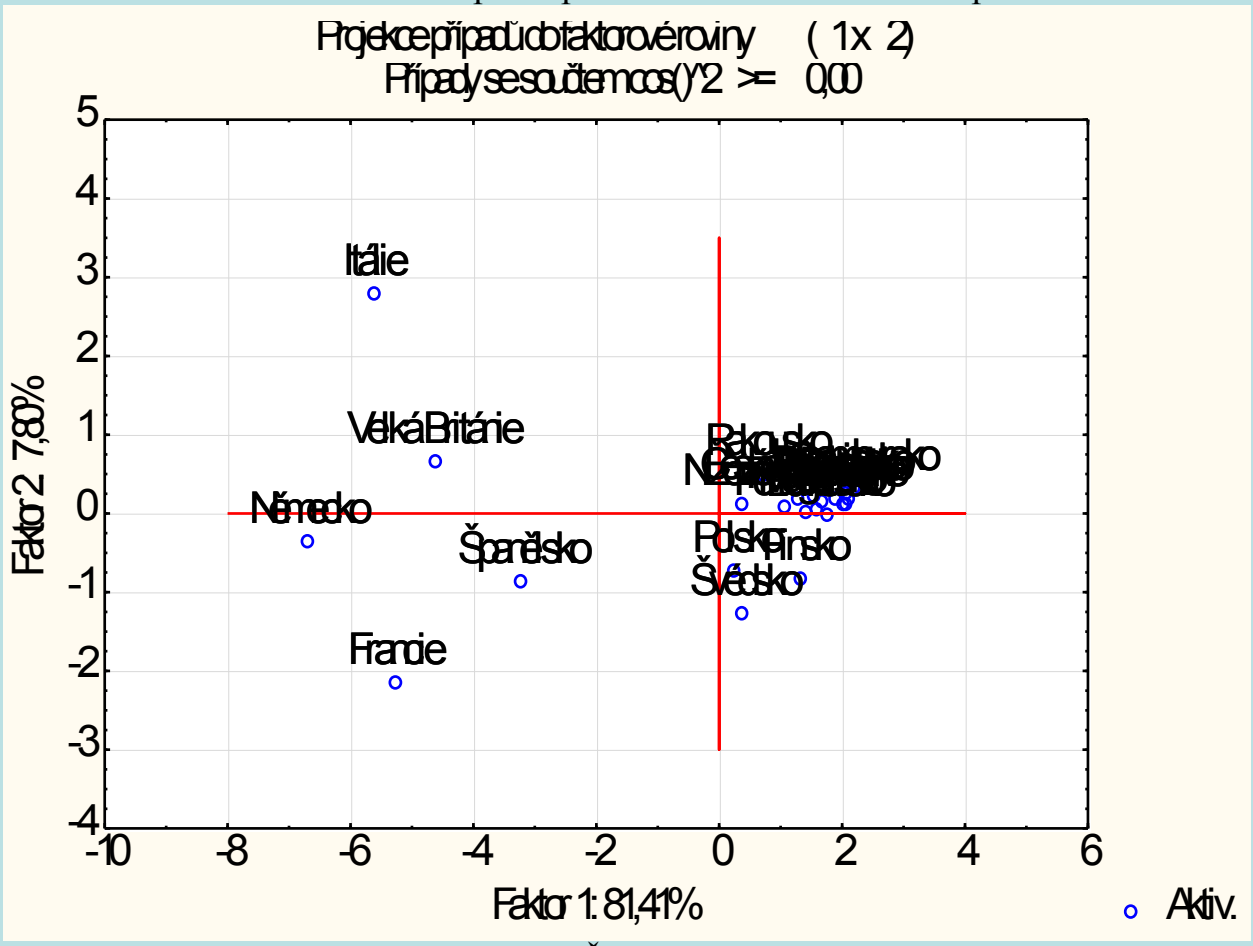
Úkol: najít skupiny států, které mají podobné podmínky na rozvoj turistického ruchu.

Krabicové diagramy jednotlivých proměnných:



Velmi rozdílná variabilita, použijeme standardizovaná data.

Znázornění rozmístění států na ploše prvních dvou hlavních komponent:



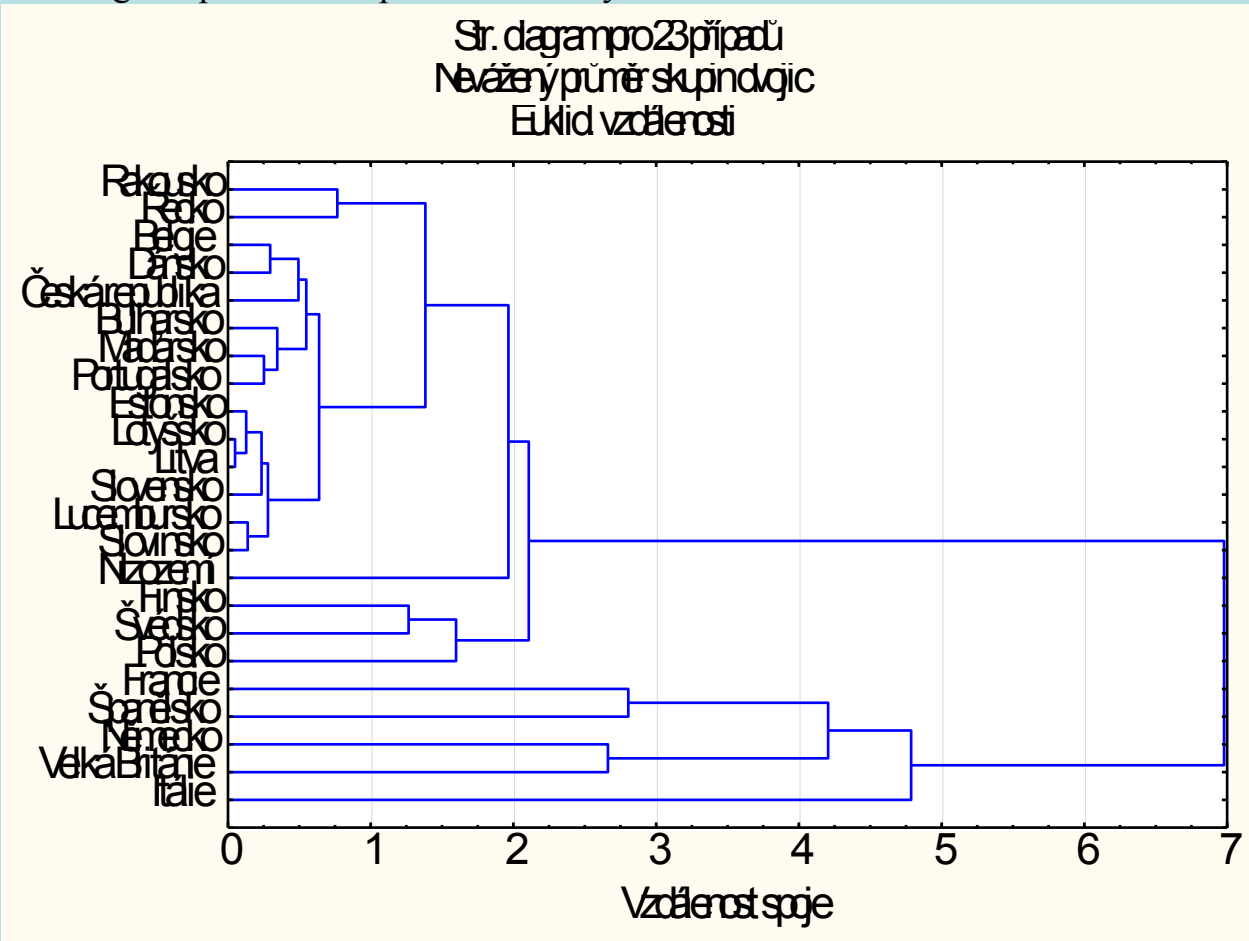
Státy Itálie, Velká Británie, Německo, Španělsko, Francie budou zřejmě tvořit jeden shluk, ostatní státy druhý shluk.

S pomocí MATLABu byly vypočítány kofenetické koeficienty korelace pro 5 shlukovacích metod: metodu nejbližšího souseda, metodu nejvzdálenějšího souseda, metodu průměrné vazby, metodu vážené průměrné vazby a Wardovu metodu:

Metoda nejbližšího souseda 0,9484
Metoda nejvzdálenějšího souseda 0,9566
Metoda průměrné vazby 0,9582
Metoda vážené průměrné vazby 0,9580
Wardova metoda 0,9453

Nejvyšší kofenetický koeficient korelace dostaneme pro metodu průměrné vazby.

Dendrogram pro metodu průměrné vazby:

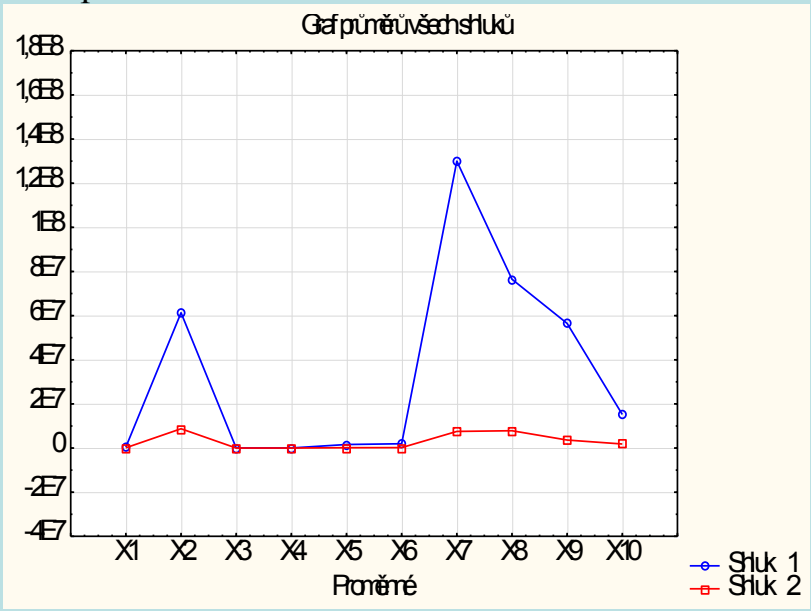


Provedeme-li řez dendrogramem na úrovni 5, získáme 2 shluky, jak bylo vidět již při znázornění rozmístění států na ploše prvních dvou hlavních komponent.

Průměry jednotlivých proměnných v 1. a 2. shluku:

Proměnná	1.shluk obsahuje 5 prip., 2. 18			
	Průměr	Směrodatná odchylka	Průměr	Směrodatná odchylka
X1	4164	1736	1141	1232
X2	61339	14092	8744	8448
X3	280	87	25	36
X4	350	354	15	18
X5	16061	3510	1747	1840
X6	19408	7458	2159	2493
X7	130026	21144	7540	62025
X8	76480	35802	7830	10778
X9	56469	16134	3625	3264
X10	15225	8240	1823	2381

Graf průměrů:



Do jednoho shluku patří státy s menší či střední rozlohou a menším počtem obyvatel, do druhého velké státy s velkým počtem obyvatel.

Ověření výsledků provedeme metodou k-průměrů pro k = 2.

Členy shluku č. 1 a vzdálenosti členů od středu shluku:

	Vzdále
Francie	0,838
Německo	0,890
Itálie	1,063
Španělsko	0,741
Velká Británi	0,634

Členy shluku č. 2 a vzdálenosti členů od středu shluku:

	Vzdále
Rakousko	0,392
Belgie	0,155
Bulharsko	0,137
Česká republ	0,103
Dánsko	0,162
Estonsko	0,229
Finsko	0,402
Řecko	0,330
Maďarsko	0,083
Lotyšsko	0,213
Litva	0,204
Lucembursko	0,281
Nizozemí	0,539
Polsko	0,575
Portugalsko	0,081
Slovensko	0,169
Slovinsko	0,242
Švédsko	0,648

Vliv jednotlivých proměnných na zařazení do shluků posoudíme ANOVOU:

Proměnná	Analýza rozptylu (turisticky_ruc					
	Mezi skupinami SČ	sv	Vnitřní SČ	sv	F	významnost p
X1	10,68	1	11,31	2	19,82	0,000
X2	18,55	1	3,442	2	113,2	0,000
X3	18,25	1	3,747	2	102,2	0,000
X4	10,22	1	11,77	2	18,24	0,000
X5	19,41	1	2,586	2	157,5	0,000
X6	17,16	1	4,838	2	74,48	0,000
X7	21,12	1	0,878	2	504,7	0,000
X8	15,88	1	6,116	2	54,52	0,000
X9	19,78	1	2,213	2	187,7	0,000
X10	14,43	1	7,567	2	40,09	0,000

Všechny proměnné jsou významné na hladině významnosti 0,05. Statistika F nabývá největší hodnoty pro X7 (počet nocí strávených v hotelech), poté pro X9 (počet příchodů do hotelů) a X5 (počet postelí v hotelech).

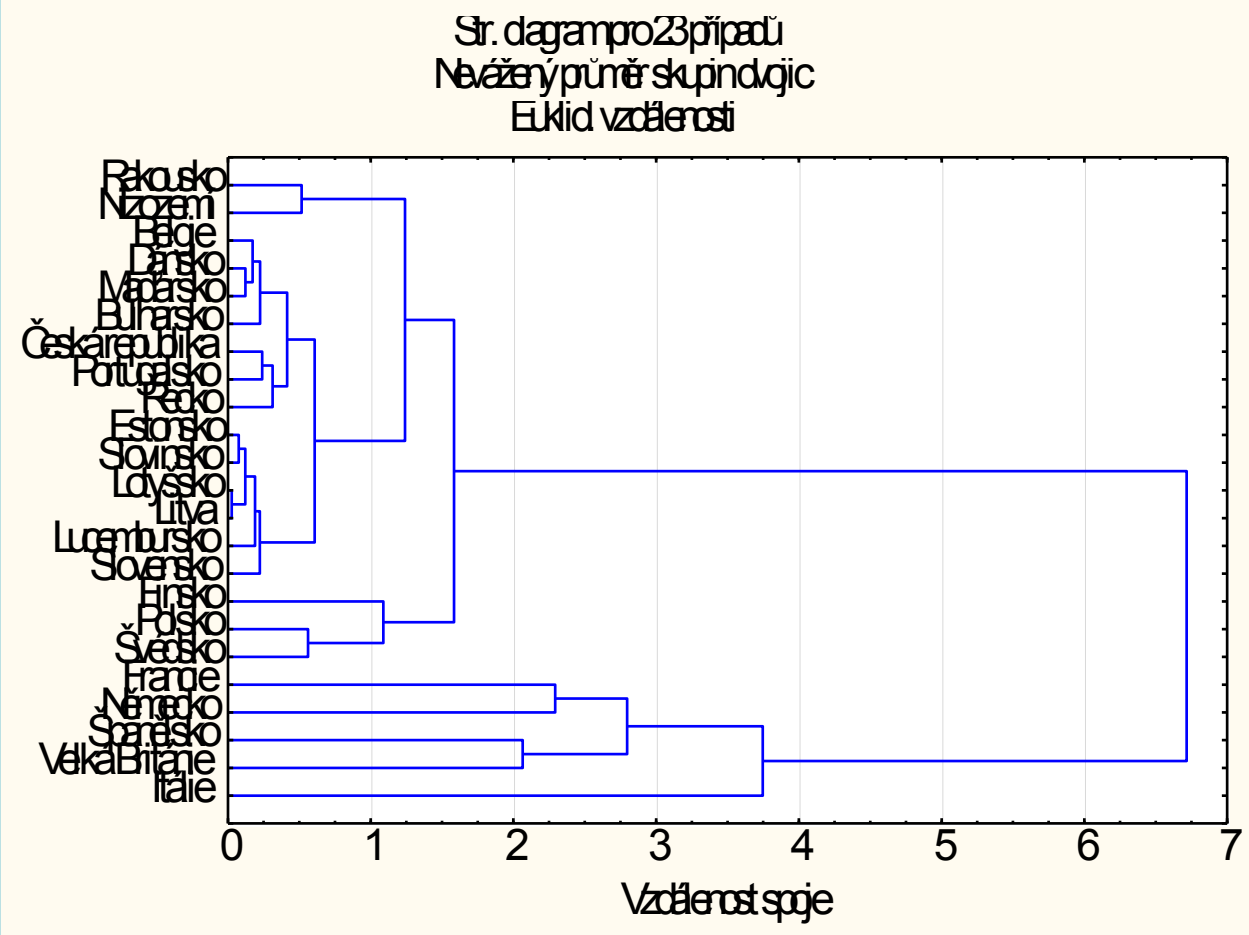
Shluková analýza provedená pomocí hlavních komponent

Použijeme první dvě hlavní komponenty. Vektory souřadnic států pro první dvě hlavní komponenty:

Případ	Fakto	Fakto
Rakousko	0,70	0,51
Belgie	1,53	0,24
Bulharsko	1,73	-0,00
Česká repu	1,26	0,21
Dánsko	1,65	0,15
Estonsko	2,07	0,19
Finsko	1,31	-0,82
Francie	-5,28	-2,13
Německo	-6,71	-0,34
Řecko	1,04	0,09
Maďarsko	1,58	0,05
Itálie	-5,60	2,79
Lotyšsko	2,05	0,12
Litva	2,02	0,11
Lucembursk	2,16	0,33
Nizozemí	0,35	0,14
Polsko	0,21	-0,71
Portugalsko	1,40	0,01
Slovensko	1,87	0,19
Slovinsko	2,07	0,27
Španělsko	-3,21	-0,84
Švédsko	0,34	-1,25
Velká Britá	-4,61	0,66

Shlukovou analýzu provedeme s proměnnými Faktor 1, Faktor 2.

Dendrogram pro metodu průměrné vazby:



Při tomto způsobu shlukování opět dostáváme stejné shluky jako v případě, kdy použijeme všech 10 proměnných.