Osnova přednášky Shluková analýza

1. 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.

1.1. Metody hledání shluků

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.

a) 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.

1.2. 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.

2. Podobnost objektů

Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost.

2.1. Matice euklidovských vzdáleností

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

$$\mathbf{D} = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$$

. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

2.2. 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	10	8
2	9	8	8	10
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 (pca)					
Případ	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0,0	3,6	12,7	12,7	12,6	14,0
P_2	3,6	0,0	12,8	13,2	12,5	14,1
P_3	12,7	12,8	0,0	2,2	3,2	4,1
P_4	12,7	13,2	2,2	0,0	3,0	3,2
P_5	12,6	12,5	3,2	3,0	0,0	2,2
P_6	14,0	14,1	4,1	3,2	2,2	0,0

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.

3.1. Algoritmus hierarchického shlukování

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.

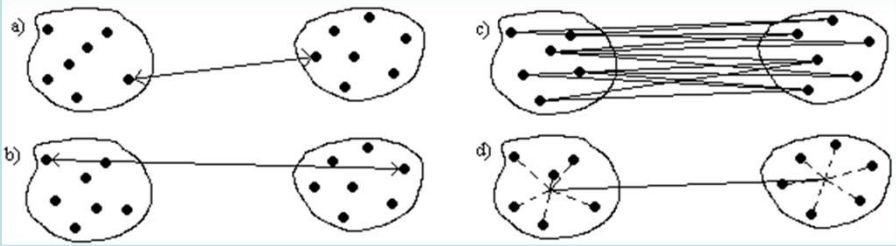
3.2. Metody počítání vzdáleností mezi shluky

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: Kritériem pro shlukování je součet kvadrátů odchylek každého objektu od těžiště shluku, do kterého náleží. Wardova metoda se hodí pro práci s objekty, které mají stejný rozměr proměnných.

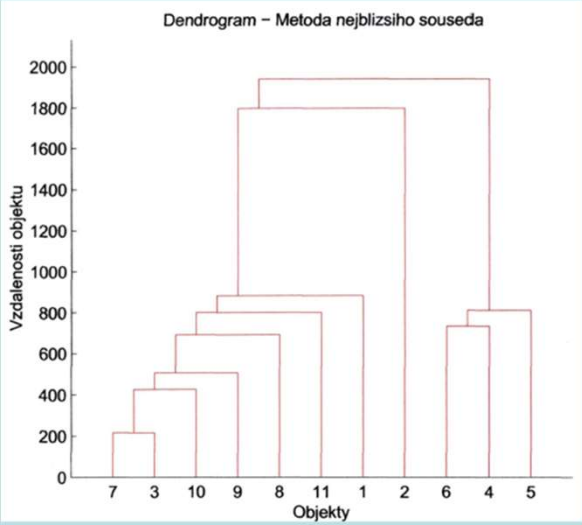


Schematické znázornění:

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.



3.3. 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: Systém STATISTICA bohužel neposkytuje 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.

3.4. 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.

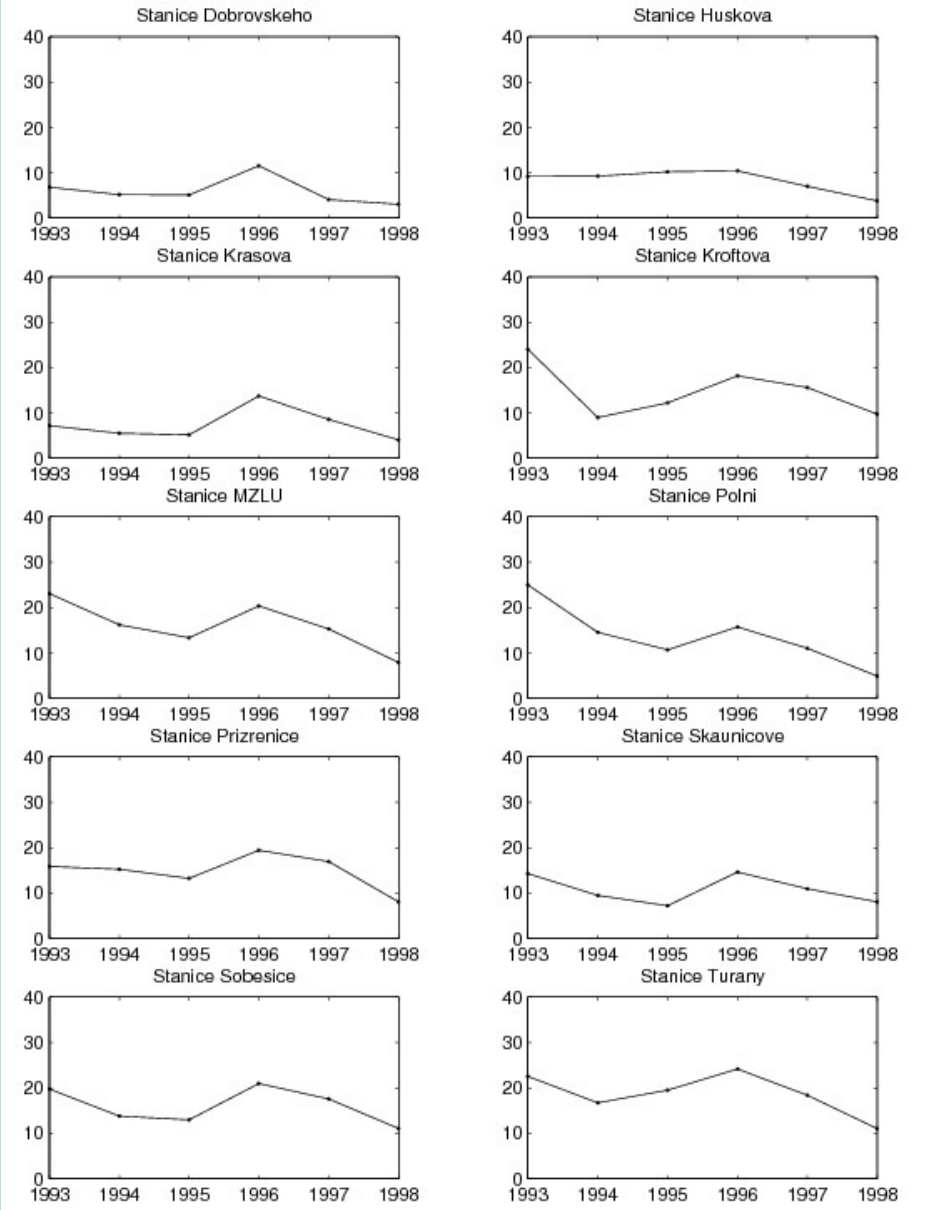
Uvedené stanice jsou obhospodařovány jednak brněnskou pobočkou ČHMÚ (to jsou stanice KRO, MZL, PRI, SOB, TUR) a jednak MHS (to jsou stanice DOB, HUS, KRA, POL, SKA). Každá z těchto organizací však zjišťuje hodnoty SO_2 jinou metodou – ČHMÚ gravimetrickou a MHS aspiračně kolorimetrickou. Teprve od r.1993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

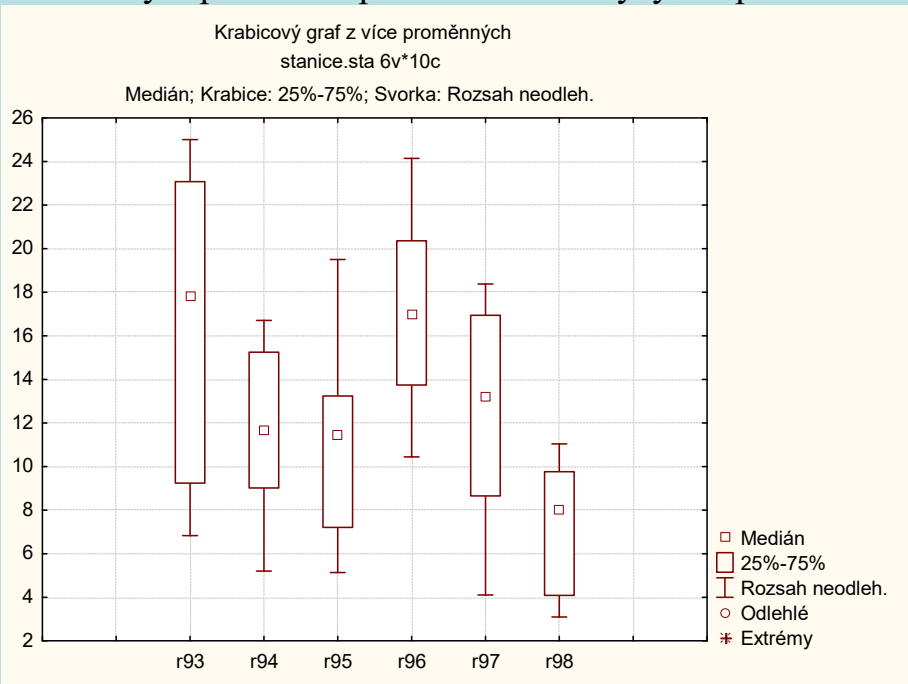
Každá ze sledovaných 10 stanic byly popsána šesti údaji, jak vidíme v této tabulce.

	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.


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

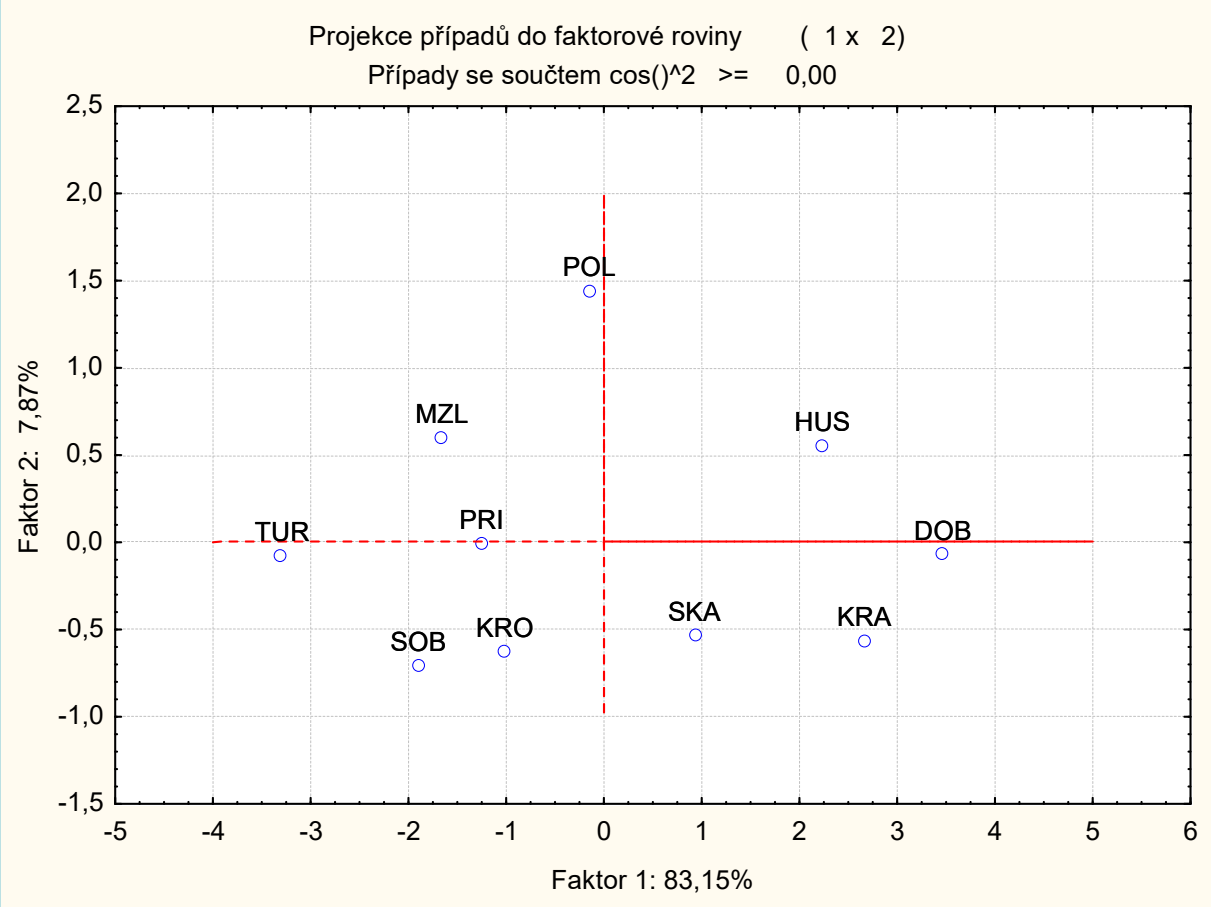
	1 r93	2 r94	3 r95	4 r96	5 r97	6 r98
DOB	-1,398	-1,457	-1,34	-1,205	-1,722	-1,363
HUS	-1,059	-0,514	-0,165	-1,459	-1,126	-1,11
KRA	-1,345	-1,38	-1,326	-0,714	-0,796	-1,034
KRO	1,0192	-0,575	0,2882	0,2906	0,619	0,8596
MZL	0,8844	1,0904	0,5441	0,7816	0,5601	0,2469
POL	1,1549	0,7081	-0,059	-0,258	-0,304	-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.

	Vlastní čísla korelační matice a související statistiky (stanice.sta) Pouze aktiv. proměnné			
Pořadí vl.č.	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%.

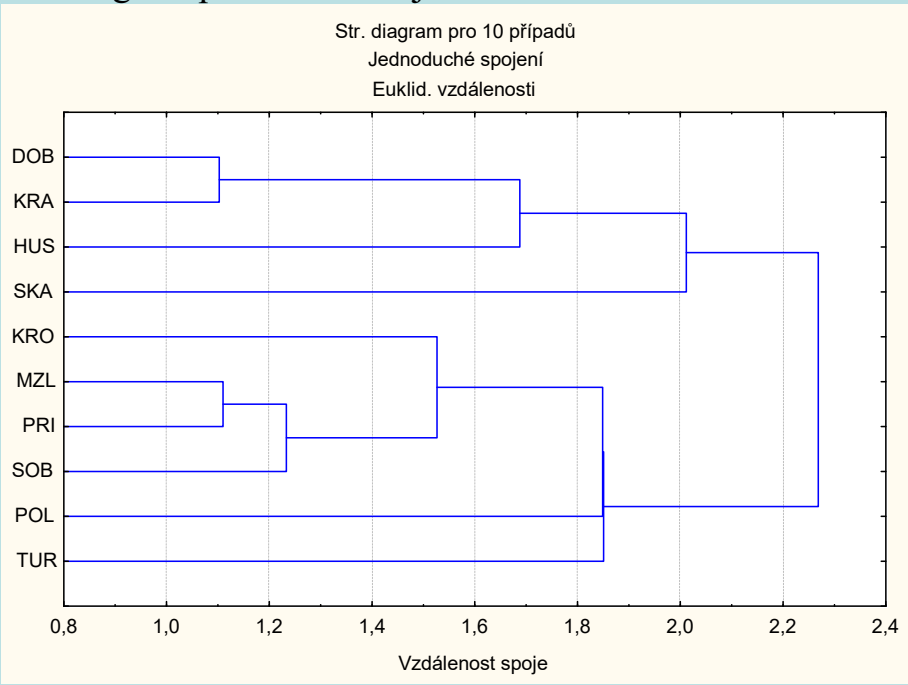
Rozmístění stanic na ploše prvních dvou hlavních komponent:



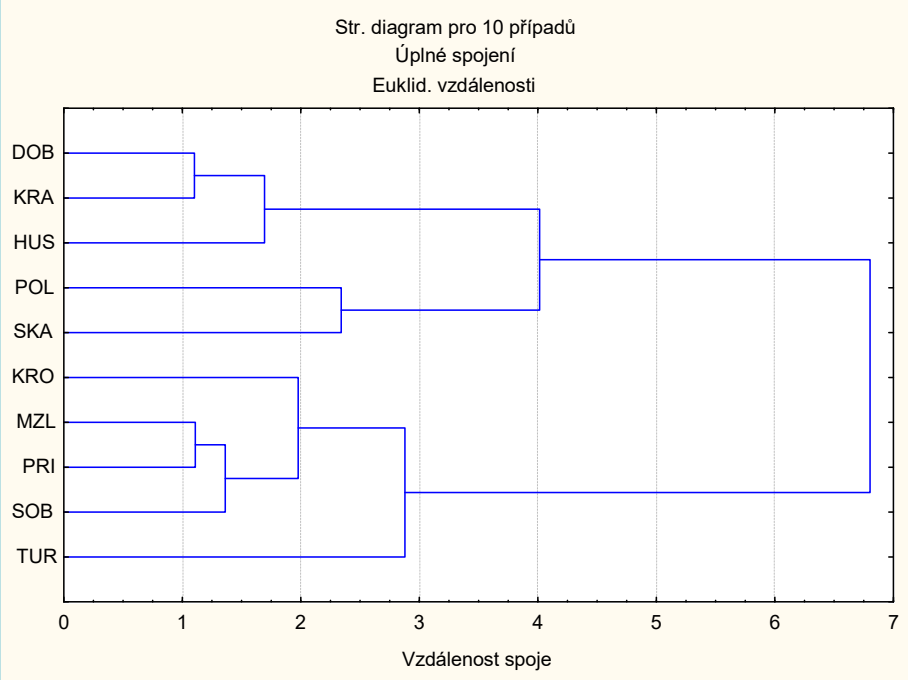
Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

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.
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.

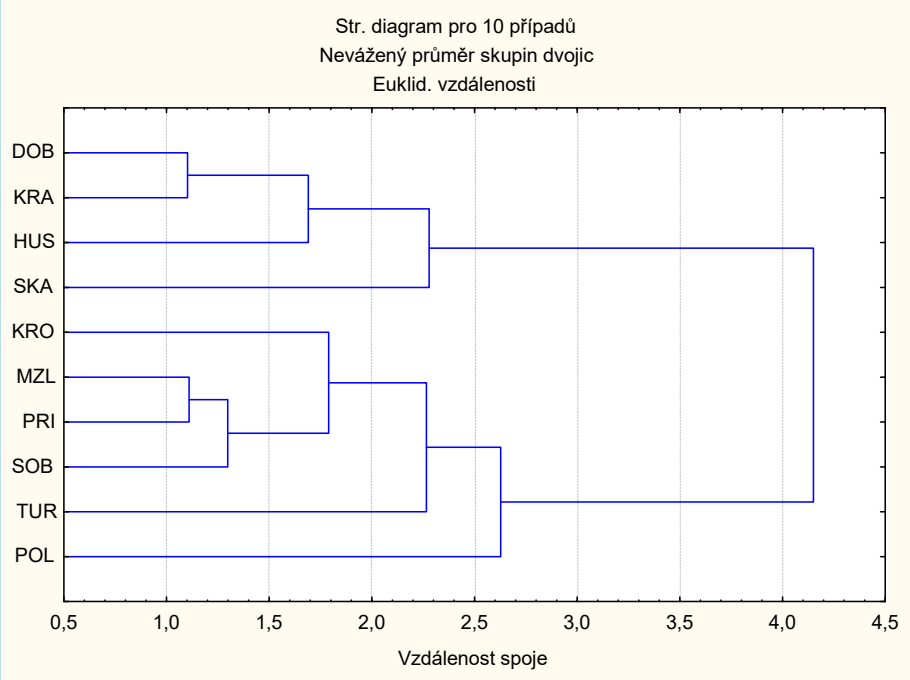
Dendrogram pro metodu nejbližšího souseda:



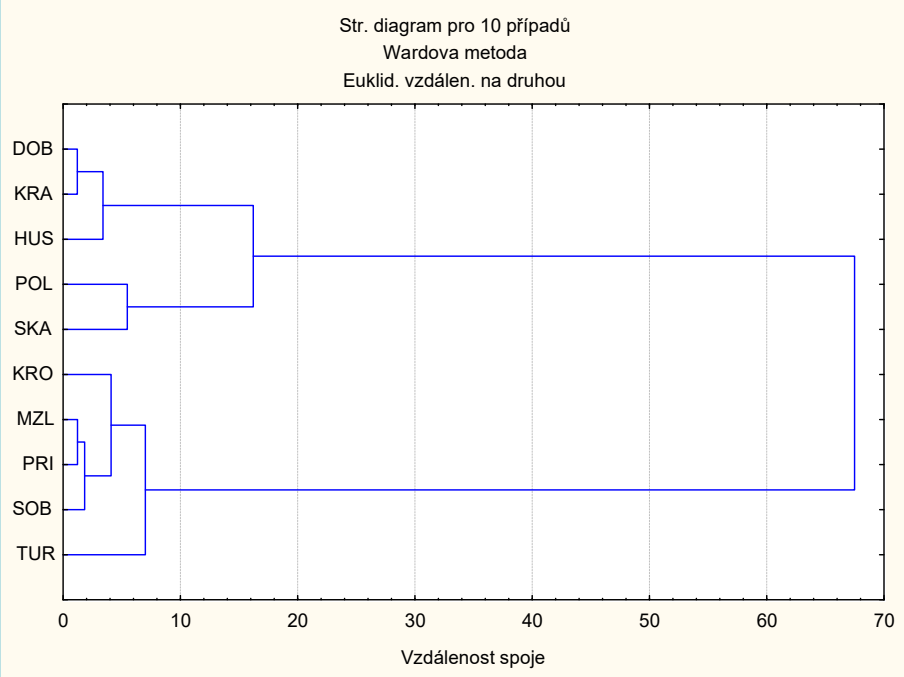
Dendrogram pro metodu nejvzdálenějšího souseda:



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	kofenetický 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:

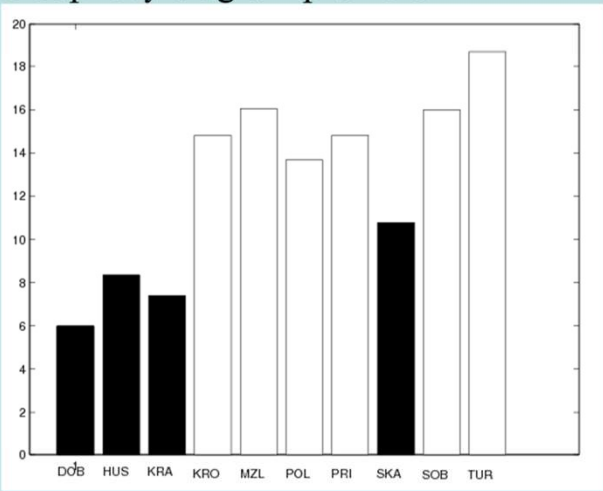
	Amalgamation Schedule (stanice.sta) Unweighted pair-group average Euclidean distances									
linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10
1,102663	DOB	KRA								
1,109963	MZL	PRI								
1,298066	MZL	PRI	SOB							
1,690602	DOB	KRA	HUS							
1,789642	KRO	MZL	PRI	SOB						
2,265795	KRO	MZL	PRI	SOB	TUR					
2,279103	DOB	KRA	HUS	SKA						
2,627296	KRO	MZL	PRI	SOB	TUR	POL				
4,150232	DOB	KRA	HUS	SKA	KRO	MZL	PRI	SOB	TUR	POL

Charakteristiky nalezených shluků

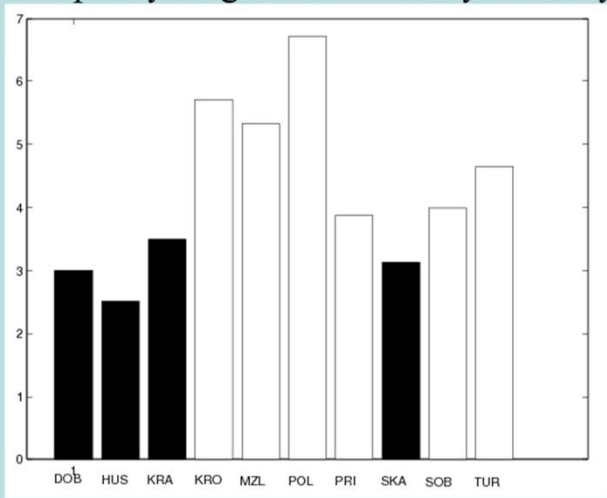
První shluk je tvořen stanicemi, které se vyznačují poměrně nízkými průměrnými ročními koncentracemi oxidu siřičitého (od 6 $\mu g/m^3$ po 11 $\mu g/m^3$ i malými směrodatnými odchylkami (od 2,5 $\mu g/m^3$ po 3,5 $\mu g/m^3$). S výjimkou stanice KRA jsou umístěny v centrální části města.

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.

4. Nehierarchické shlukování – 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:

4.1. Algoritmus metody k-průměrů

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.

4.2. Příklad

Aplikujte metodu k-průměrů na data o stanicích.
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

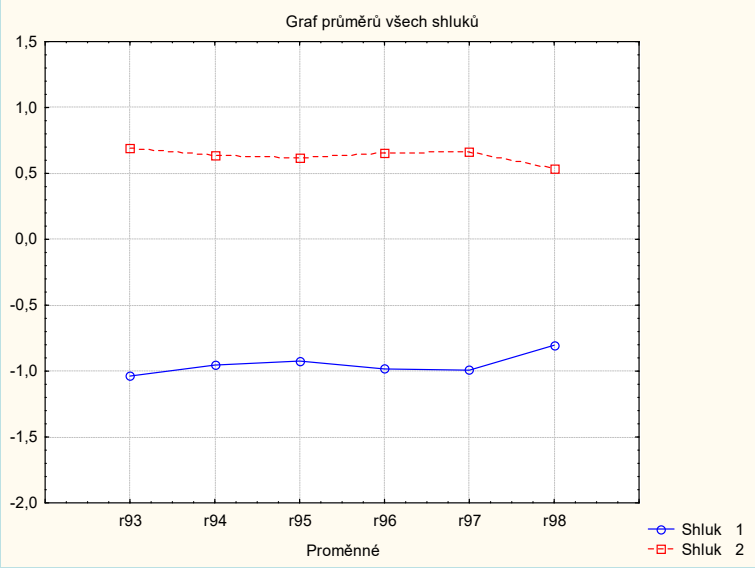
Vidíme, že metoda k průměrů dospěla k témuž výsledku jako metoda průměrné vazby.
1. shluk: DOB, KRA, HUS, SKA.
2. shluk: MZL, PRI, SOB, KRO, TUR, POL.

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)					
	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.

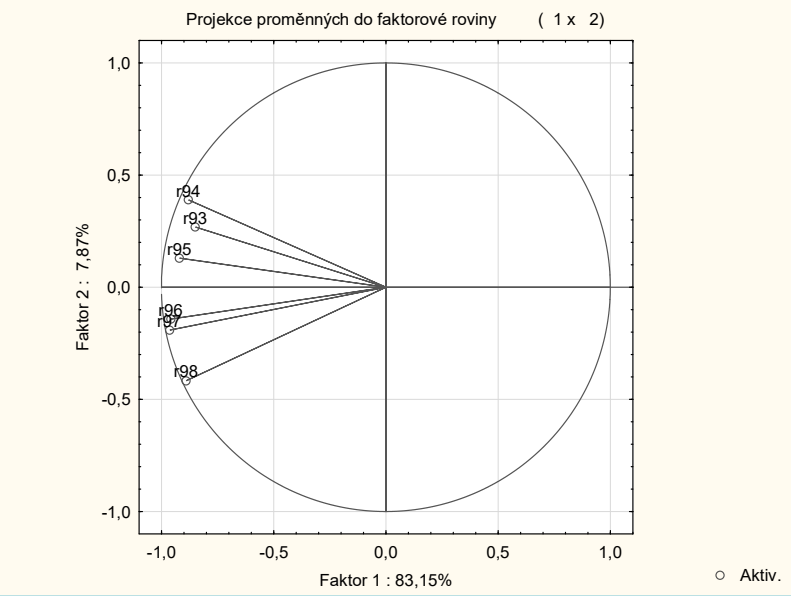
Graf průměrů obou shluků



5. Shlukování proměnných

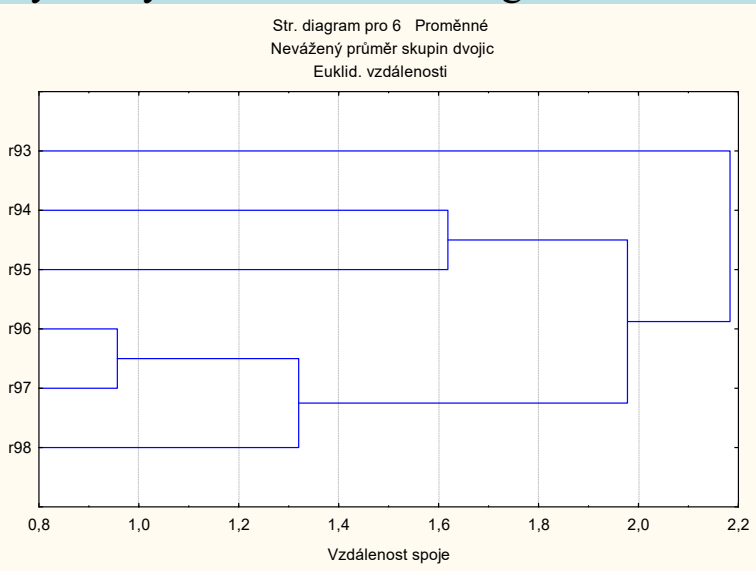
Systém STATISTICA pomocí shlukové analýzy umožňuje zjistit, které proměnné mají k sobě blízko.
Budeme pracovat se standardizovanými hodnotami datového souboru stanice.sta.

Znázorníme jednotlivé roky na ploše prvních dvou hlavních komponent:



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

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 (stanice.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 3 příp.
	Vzdálen.
r96	0,231170
r97	0,162806
r98	0,261356

	Členy shluku číslo 2 (stanice.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 2 příp.
	Vzdálen.
r94	0,255911
r95	0,255911

	Členy shluku číslo 3 (stanice.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 1 příp.
	Vzdálen.
r93	0,00

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

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

Proměnná	Analýza rozptylu (stanice.sta)					
	Mezisk. SČ	sv	Vnitřní SČ	sv	F	význam. p
DOB	0,001536	2	0,147511	3	0,01562	0,984583
HUS	0,982796	2	0,138590	3	10,63708	0,043448
KRA	0,378373	2	0,056645	3	10,01957	0,046988
SKA	0,264229	2	0,466157	3	0,85024	0,509882
KRO	1,080078	2	0,535548	3	3,02516	0,190847
MZL	0,147265	2	0,293626	3	0,75231	0,543493
POL	2,085189	2	0,446093	3	7,01150	0,073982
PRI	0,490681	2	0,236837	3	3,10771	0,185741
SOB	0,561752	2	0,076213	3	11,05624	0,041290
TUR	0,400365	2	0,395111	3	1,51994	0,350057

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

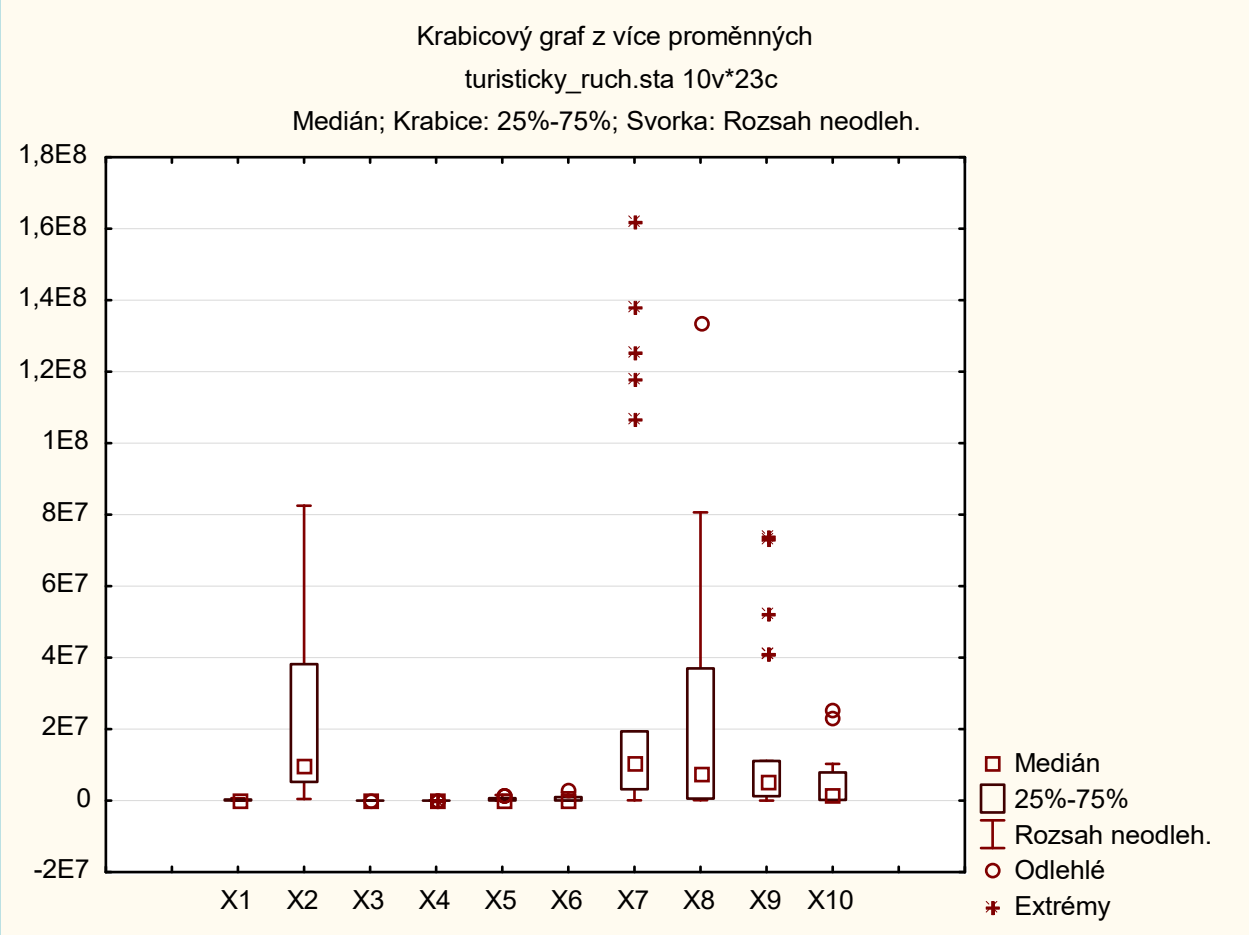
6. Analýza turistického ruchu ve 23 státech EU

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	30528	10445852	1899	1550	121000	295000	4313000	8514000	2364000	2208000
Bulharsko	110910	7761049	1230	325	201000	20000	3957000	490000	1721000	173000
Česká republika	78866	10220577	4278	3327	232000	201000	8601000	12124000	3388000	2637000
Dánsko	43094	5411405	482	608	70000	323000	5316000	11556000	1899000	1624000
Estonsko	45226	1347510	317	467	25000	13000	751000	378000	428000	191000
Finsko	338145	5236611	938	459	118000	93000	10388000	2372000	5948000	1061000
Francie	674843	62637596	19811	9244	1740000	3039000	125216000	62426000	73066000	10291000
Itálie	301318	58462375	33527	96409	2028000	2322000	138222000	68504000	41295000	8918000
Litva	65200	3425324	331	193	20000	11000	728000	494000	347000	158000
Lotyšsko	64589	2306434	337	81	19000	5000	796000	225000	354000	71000
Lucembursko	2586,4	461230	293	252	14000	52000	85000	145000	29000	34000
Maďarsko	93030	10097549	2061	1056	162000	167000	6622000	2336000	2778000	839000
Německo	357021	82500849	36593	18756	1621000	1696000	161895000	133840000	73777000	25296000
Nizozemí	41526	16305526	3135	4025	192000	998000	14375000	40575000	8301000	7881000
Polsko	312679	38173835	2200	4523	170000	400000	12464000	25612000	6805000	5482000
Portugalsko	92345	10529255	2012	288	264000	183000	11648000	6230000	5274000	1214000
Rakousko	83872	8206524	14267	6281	571000	355000	19383000	7915000	6896000	1532000
Řecko	131990	11082751	9036	341	682000	96000	13942000	587000	5933000	131000
Slovensko	49035	5384822	885	1131	57000	103000	3183000	2638000	1244000	656000
Slovinsko	20273	1997590	344	358	30000	35000	1653000	1405000	459000	353000
Španělsko	504030	43038035	17607	17151	1580000	1484000	106875000	36999000	41600000	8552000
Švédsko	449964	9011392	1857	2089	197000	537000	17518000	17345000	11096000	6586000
Velká Británie	244820	60059900	32926	33877	1062000	1163000	117926000	80635000	52611000	23069000

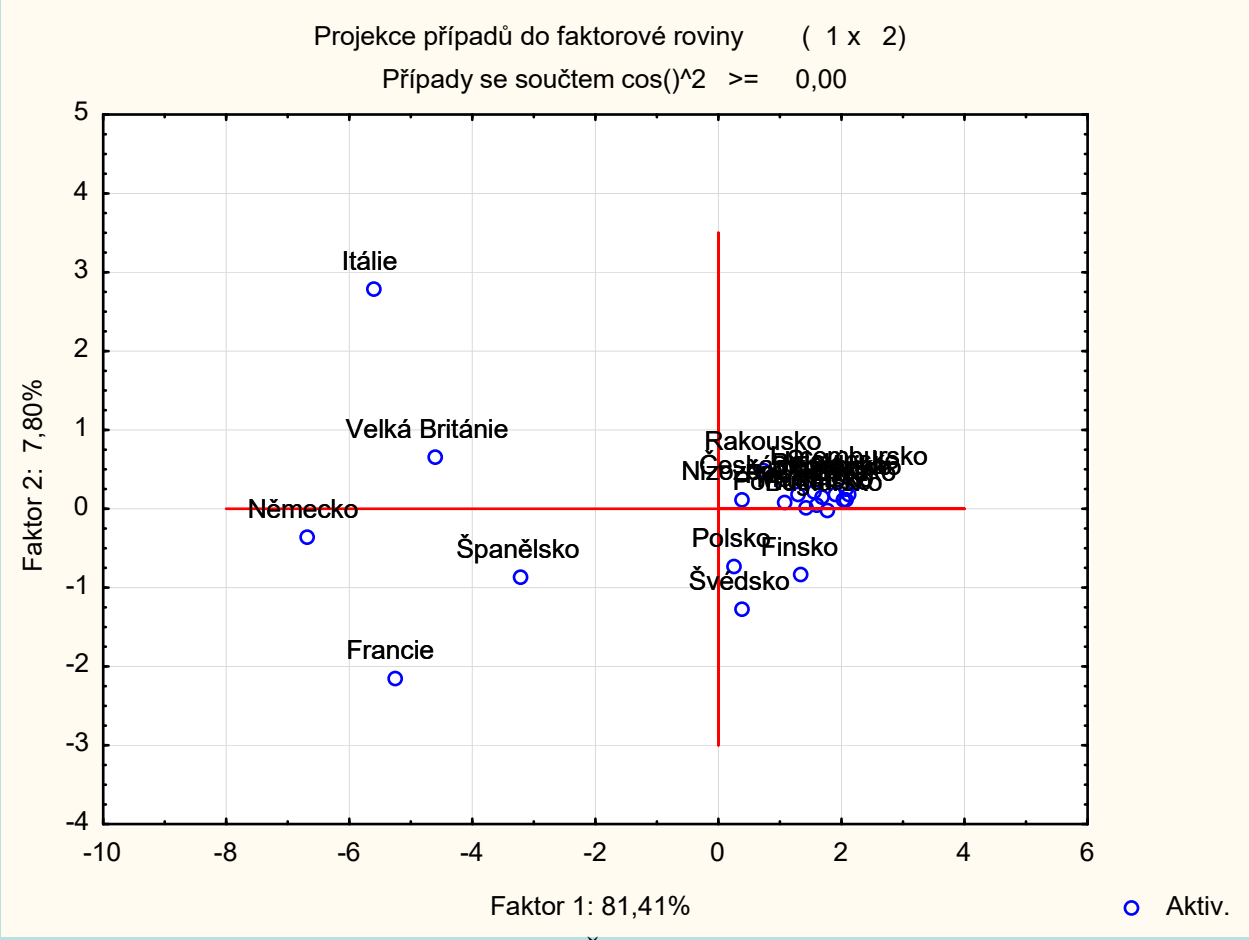
X1 … rozloha, X2 … počet obyvatel, X3 … počet hotelů, X4 … počet jiných ubytovacích zařízení, X5 resp. X6 … počet postelí v hotelech resp. jiných ubytovacích zařízeních, X7 resp. X8 … počet nocí strávených v hotelech resp. jiných ubytovacích zařízeních, X9 resp. X10 … počet příchodů do hotelů resp. jiných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



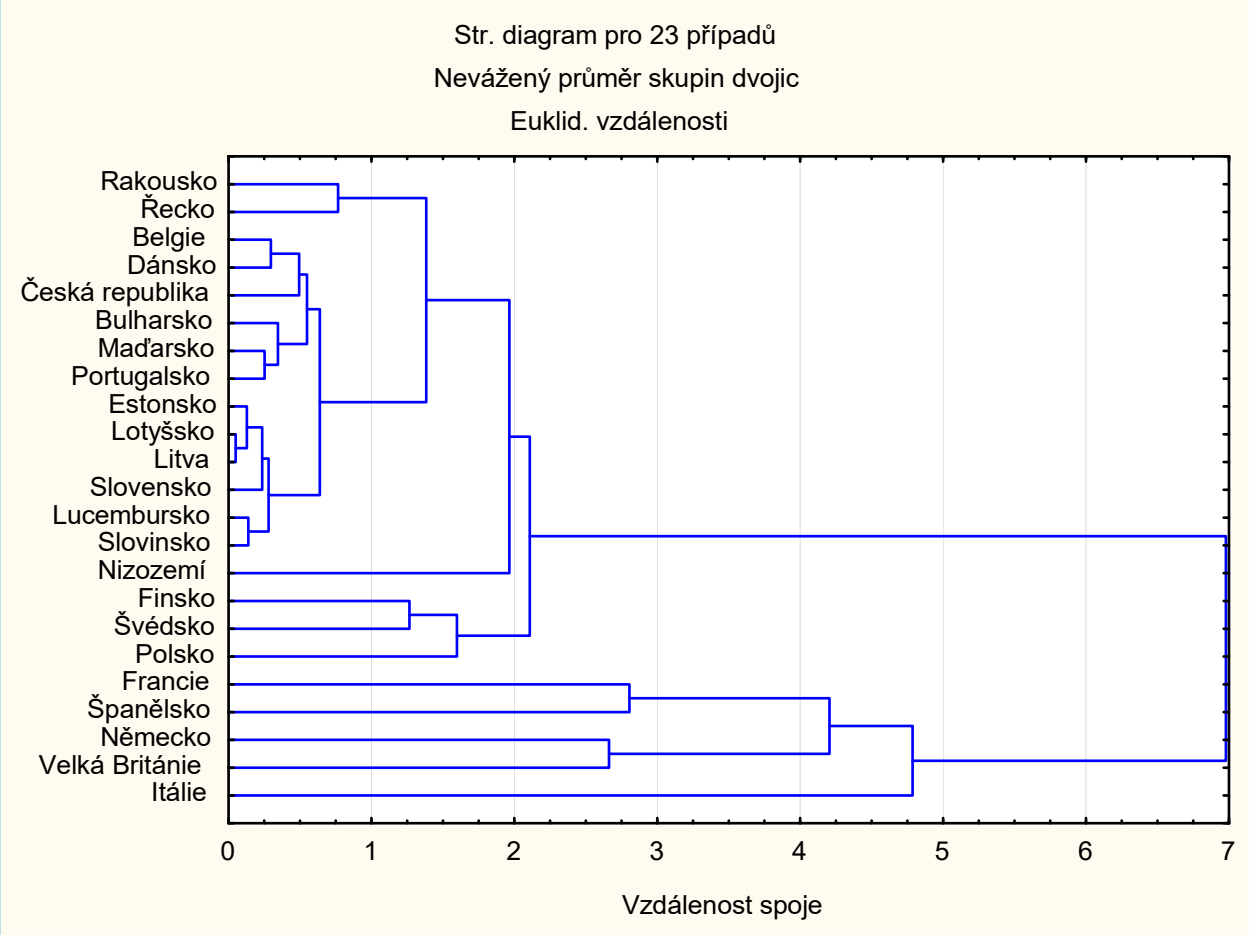
Státy Itálie, Velká Británie, Německo, Španělsko, Francie budou zřejmě tvořit jeden shluk, ostatní státy druhý shluk.

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:

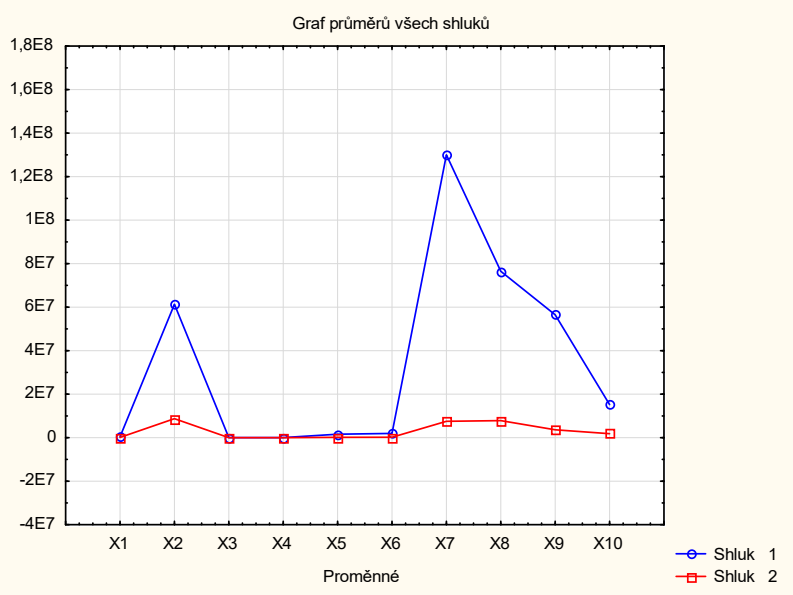


Provedeme-li řez dendrogramem na úrovni 5, získáme 2 shluky, jak bylo vidět již při znázornění rozmístění států na ploše prvních dvou hlavních komponent.

Průměry jednotlivých proměnných v 1. a 2. shluku:

Proměnná	1.shluk obsahuje 5 příp., 2. 18 příp.			
	Průměr 1	Směrod. odchylka 1	Průměr 2	Směrod. odchylka 2
X1	416406	173696	114103	123238
X2	6133975(	1409208(	8744735	8448575
X3	28093	8713	2550	3601
X4	35087	35419	1520	1817
X5	160620(	351026	174722	184007
X6	194080(	745809	215944	249384
X7	13002680(	21144280	7540167	6202994
X8	7648080(	3580264(	783005(	1077884(
X9	5646980(	1613419(	3625778	3264718
X10	1522520(	8240220	1823945	2381825

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:

	Vzdálen.
Francie	0,838833
Německo	0,890074
Itálie	1,063136
Španělsko	0,741349
Velká Británie	0,634043

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

	Vzdálen.
Rakousko	0,392711
Belgie	0,155856
Bulharsko	0,137693
Česká republika	0,103918
Dánsko	0,162968
Estonsko	0,229445
Finsko	0,402751
Řecko	0,330998
Maďarsko	0,083707
Lotyšsko	0,213803
Litva	0,204610
Lucembursko	0,281346
Nizozemí	0,539286
Polsko	0,575224
Portugalsko	0,081809
Slovensko	0,169324
Slovinsko	0,242607
Švédsko	0,648680

Vliv jednotlivých proměnných na zařazení do shluků posoudíme ANOVOU:

	Analýza rozptylu (turisticky_ruch.sta)					
Proměnná	Mezisk. SČ	sv	Vnitřní SČ	sv	F	význam. p
X1	10,68234	1	11,31766	21	19,8212	0,000220
X2	18,55778	1	3,44222	21	113,2158	0,000000
X3	18,25299	1	3,74701	21	102,2983	0,000000
X4	10,22861	1	11,77139	21	18,2477	0,000339
X5	19,41312	1	2,58688	21	157,5938	0,000000
X6	17,16160	1	4,83840	21	74,4860	0,000000
X7	21,12128	1	0,87872	21	504,7651	0,000000
X8	15,88304	1	6,11696	21	54,5277	0,000000
X9	19,78645	1	2,21355	21	187,7144	0,000000
X10	14,43836	1	7,56164	21	40,0979	0,000003

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).

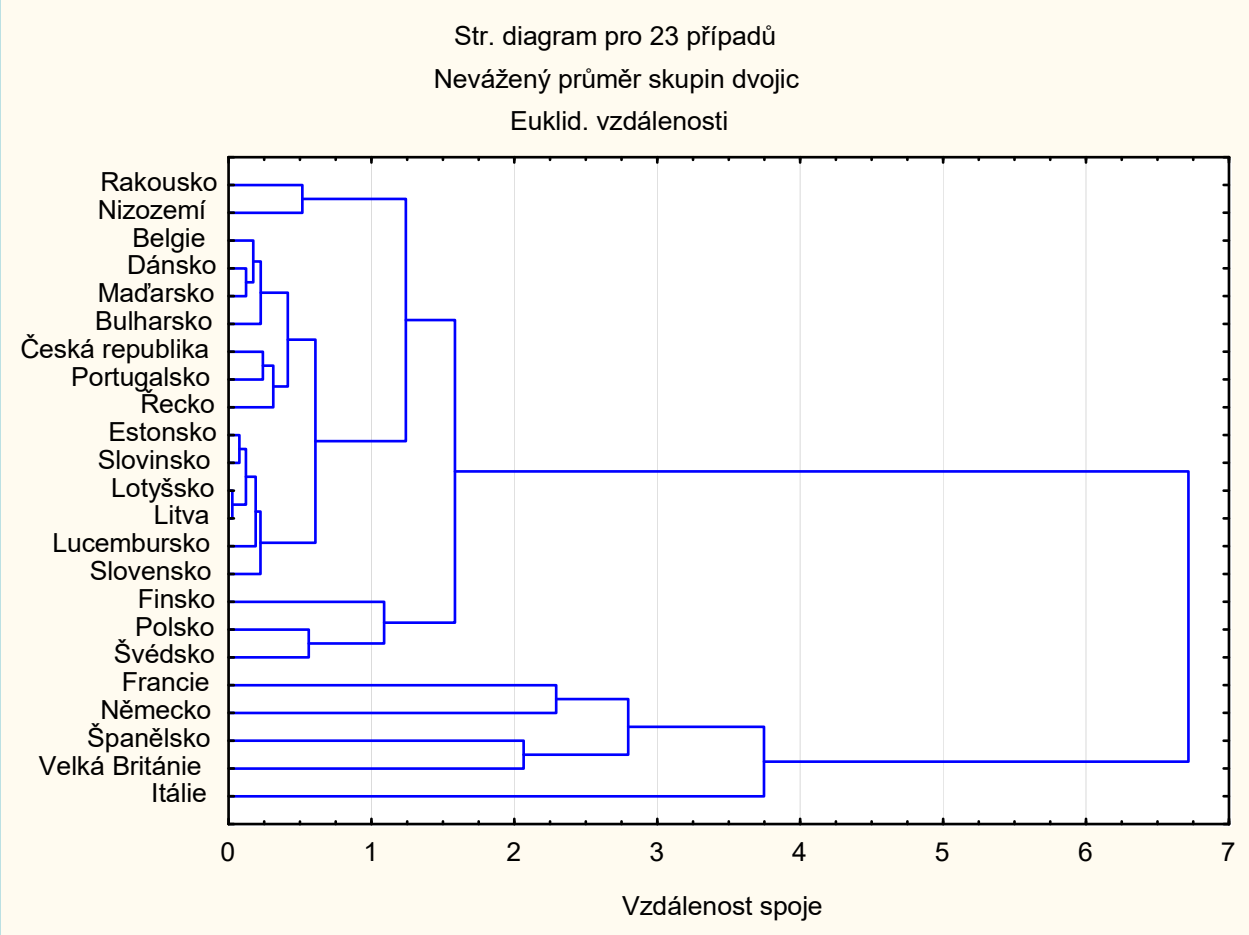
Shluková analýza provedená pomocí hlavních komponent

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

Případ	Faktor 1	Faktor 2
Rakousko	0,70756	0,51743
Belgie	1,53638	0,24411
Bulharsko	1,73974	-0,00915
Česká republika	1,26767	0,21195
Dánsko	1,65269	0,15232
Estonsko	2,07905	0,19891
Finsko	1,31958	-0,82216
Francie	-5,28382	-2,13623
Německo	-6,71886	-0,34872
Řecko	1,04156	0,09725
Maďarsko	1,58274	0,05162
Itálie	-5,60234	2,79753
Lotyšsko	2,05626	0,12158
Litva	2,02975	0,11947
Lucembursko	2,16385	0,33312
Nizozemí	0,35460	0,14050
Polsko	0,21670	-0,71265
Portugalsko	1,40602	0,01546
Slovensko	1,87153	0,19528
Slovinsko	2,07055	0,27464
Španělsko	-3,21569	-0,84866
Švédsko	0,34390	-1,25937
Velká Británie	-4,61943	0,66578

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.