APLIKACE SHLUKOVÉ ANALÝZY V EKOLOGII

MARIE BUDÍKOVÁ

Abstrakt. In this paper, the basic principles of hierarchical cluster analysis are described. An example of calculation and application of cluster analysis is presented. This example illustrates how the different regions in Brno can by classified by means of annual concentrations of SO_2.

Резюме: Эта работа показывает главные принципы иерархической группировки объектоб. Показан один пример этой группировки. Станции в городе Брно вводятся в группы на основании концентраций SO_2.

1. Vymezení problému

Na katedře geografie Přírodovědecké fakulty Masarykovy univerzity v Brně byly v rámci bakalářské práce [5] shromážděny údaje o průměrných měsíčních hodnotách oxidu siřičitého v období 1984 - 1998 na deseti monitorovacích stanicích 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, TUR. Rozmístění stanic na území Brna znázorňuje obr. 1

Obr. 1 - Rozmístění monitorovacích stanic na území Brna

2000 *Mathematics Subject Classification.* Primary 62H30; Secondary 62P12.

Klíčová slova. Aglomerativní hierarchické shlukování, dendrogram, kofenetický koeficient korelace, fúzní koeficient.

Příspěvek vznikl s podporou výzkumného záměru MŠMT MSM 143100001.

Tyto údaje by měly - mimo jiné - posloužit k řešení problému optimalizace sítě monitorovacích stanic. Cílem tedy bylo najít stanice, které mají podobné rysy chování. K dosažení tohoto cíle byly použity postupy shlukové analýzy. Výpočty byly provedeny pomocí systému SPSS 8.0.

Uvedené stanice jsou obhospodařovány jednak brněnskou pobočkou Českého hydrometeorologického ústavu (to jsou stanice KRO, MZL, PRI, SOB, TUR) a jednak Městskou hygienickou stanicí (DOB, HUS, KRA, POL, SKA). Každá z těchto institucí 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 tedy byly zahrnuty údaje až od r. 1993. Zabývali jsme se průměrnými ročními koncentracemi SO_2. Podle zákona o ochraně ovzduší před znečišťujícími látkami činí nejvyšší přípustná průměrná roční koncentrace SO_2 60 μg/m^3.

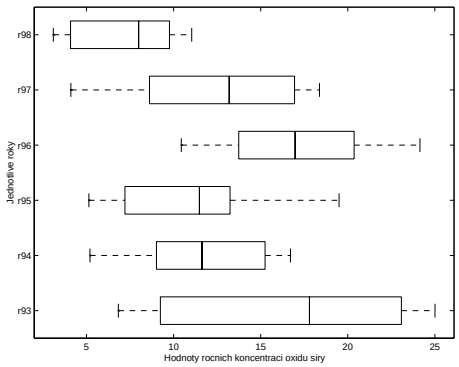
2. Popis a úprava dat

Tabulka 1 obsahuje hodnoty průměrných ročních koncentrací SO_2 v μg/m^3 v letech 1993 - 1998 na sledovaných deseti stanicích.

	Rok pozorování					
Stanice	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**
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	15.760	11.068	4.916
PRI	15.874	15.251	13.241	19.435	16.943	8.081
SKA	14.297	9.490	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

Tabulka 1 - Průměrné roční koncentrace SO_2

Prvním krokem při zpracování naměřených koncentrací bylo provedení průzkumové analýzy dat pomocí krabicových diagramů.



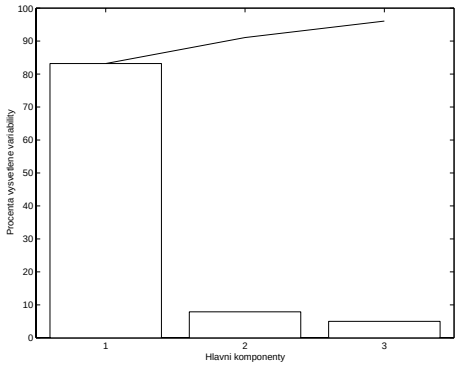
Obr. 2 - Krabicové diagramy ročních koncentrací oxidu siřičitého

Z obr. 2 je zřejmé, že údaje v jednotlivých letech vykazují dosti rozdílnou variabilitu, největší v r. 1993, nejmenší v r. 1998. Přistoupili jsme tedy ke standardizaci a nadále pracovali se standardizovanými hodnotami.

3. Metoda hlavních komponent

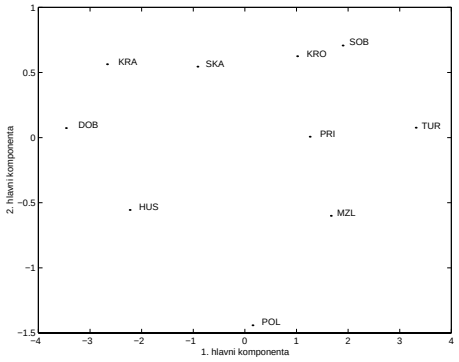
Při grafickém znázornění vícerozměrných dat se velmi často používá metoda hlavních komponent, která informace obsažené v datech dokáže vyjádřit několika málo novými proměnnými, které jsou získány jako lineární kombinace původních proměnných. Nazývají se hlavní komponenty.

Z Paretova diagramu hlavních komponent sestrojeného pro naše data vyplývá, že první hlavní komponenta vyčerpává asi 83 % variability obsažené v datech, druhá asi 8 % a třetí pouhých 5 %.



Obr. 3 - Paretův diagram hlavních komponent

Rozmístění stanic na ploše prvních dvou hlavních komponent je znázorněno na obr. č. 4.



Obr. 4 - Poloha stanic v novém souřadném systému

Z obr. 4 bychom mohli usoudit, že stanice DOB, KRA, HUS, SKA tvoří jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

4. Shluková analýza

4.1. **Cíl shlukové analýzy.** Cílem shlukové analýzy je roztřídění n objektů, z nichž každý je popsán p-rozměrným vektorem pozorování, do několika pokud možno homogenních shluků. Přesný počet shluků většinou není předem znám. Požadujeme, aby objekty uvnitř jednotlivých shluků si byly podobné co nejvíce, zatímco objekty z různých shluků si mají být podobné co nejméně.

4.2. **Podobnost objektů.** Podobnost či rozdílnost objektů posuzujeme pomocí různých měr. Pro proměnné intervalového či poměrového typu se nejčastěji používá euklidovská vzdálenost. Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

4.3. **Hierarchické shlukování.** Při aplikacích shlukové analýzy se velmi často používá aglomerativní hierarchická procedura. Její princip spočívá (viz [4]) 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.

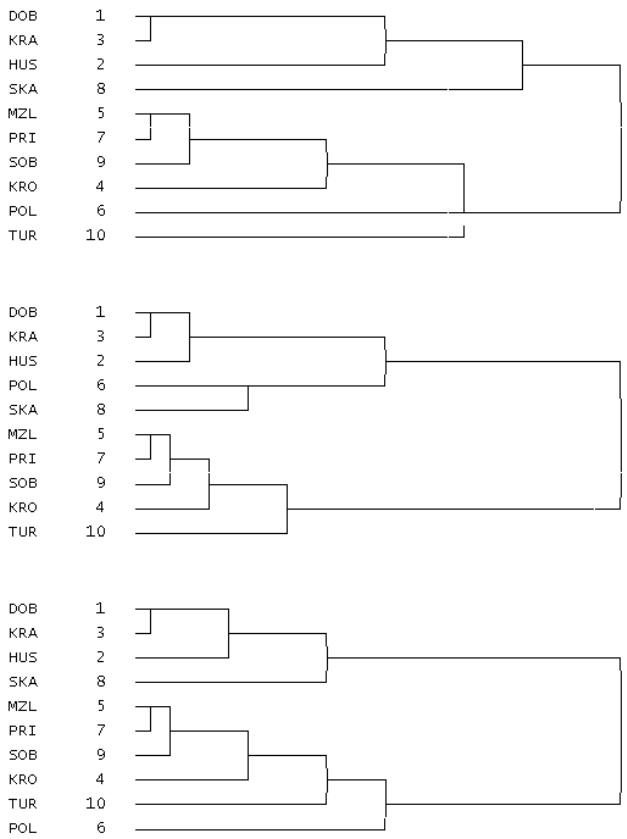
Funkce algoritmu končí, až jsou všechny objekty spojeny do jediného shluku.

Vzdálenost mezi shluky se počítá různými způsoby. V našem případě jsme použili tři metody, a to metodu nejbližšího souseda, nejvzdálenějšího souseda a metodu průměrné vazby.

Popis metod:
Metoda nejbližšího souseda: Vzdálenost mezi dvěma shluky je minimem ze všech vzdáleností mezi jejich objekty.
Metoda nejvzdálenějšího souseda: Vzdálenost mezi dvěma shluky je maximem ze všech vzdáleností mezi jejich objekty.
Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.

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

Na obr. 5 jsou zachyceny výsledky uvedených tří shlukovacích procedur.



Obr. 5 - Dendrogramy pro metodu nejbližšího souseda, nejvzdálenějšího souseda a metodu průměrné vazby

4.4. **Kofenetický koeficient korelace.** Jak je vidět na obr. 5, mohou různé shlukovací procedury poskytovat různé výsledky. K posouzení shody mezi maticí vzdáleností objektů a dendrogramem lze použít např. kofenetický koeficient korelace (viz [3]). 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. Z uvažovaných shlukovacích metod pak vybereme tu, která poskytuje nejvyšší kofenetický koeficient korelace.

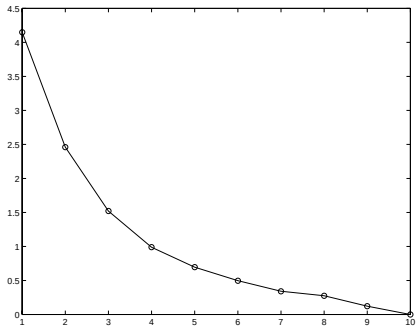
Hodnoty kofenetických koeficientů:

- metoda nejbližšího souseda: 0,737
- metoda nejvzdálenějšího souseda: 0,715
- metoda průměrné vazby: 0,783.

Nejvyšší kofenetický koeficient korelace byl dosažen pro metodu průměrné vazby. Budeme ji tedy nadále používat.

4.5. **Stanovení optimálního počtu shluků.** Pro řešení tohoto problému byla vypracována celá řada postupů - od heuristických až po formální testy (viz [2]). Z heuristických metod zde uvedeme graf závislosti hodnot fúzních koeficientů na počtu shluků (viz [1]). Jedná se o analogii tzv. „scree testu“ známého z faktorové analýzy.

Fúzním koeficientem pro k shluků rozumíme průměr maximálních vzdáleností uvnitř těchto k shluků. Pokud se v grafu závislosti hodnot fúzních koeficientů na počtu shluků objeví určité zploštění, svědčí to o tom, že zmenšení počtu shluků již není vhodné.



Obr. 6 - Průběh závislosti hodnot fúzních koeficientů na počtu shluků

Z obr. 6 vyplývá, že bude vhodné rozdělit stanice do dvou shluků. Při pohledu na dendrogram pro metodu průměrné vazby zjistíme, že 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.

4.6. **Metoda k-průměrů.** Chceme-li porovnat výsledek dané hierarchické shlukovací metody s jiným postupem, můžeme tak učinit např. pomocí metody k-průměrů, což je nehierarchická shlukovací metoda vycházející z následujícího algoritmu popsaného např. v [4]:

Algoritmus:

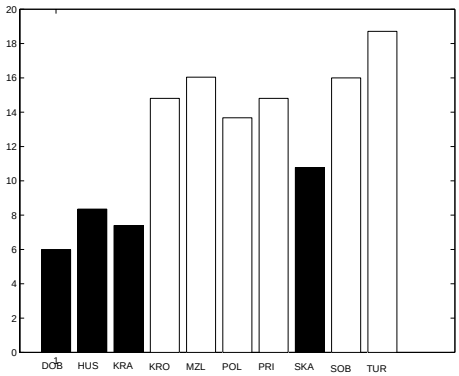
1. **krok:** Stanovíme počáteční rozklad množiny n objektů do k shluků. Rozklad zpravidla volíme náhodně.
2. **krok:** Určíme výběrové centroidy v aktuálních shlucích. (Výběrovým centroidem shluku rozumíme hypotetický objekt, jehož vektor pozorování je roven vektoru výběrových průměrů počítaných pro všechny objekty patřící 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.

V našem případě pro k=2 dospěla metoda k-průměrů po dvou iteracích 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.

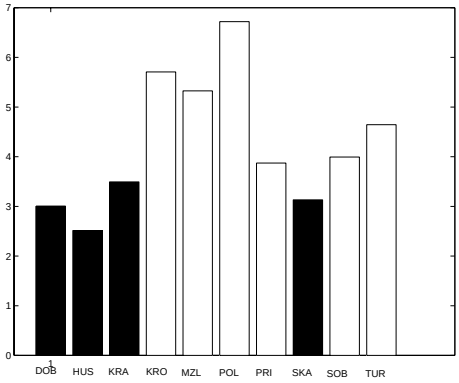
5. Charakteristiky nalezených shluků

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).



Obr. 7 - Průměry průměrných ročních koncentrací SO_2 pro 1. a 2. shluk. (Černě - 1. shluk, bíle - 2. shluk)



Obr. 8 - Směrodatné odchylky průměrů průměrných ročních koncentrací SO_2 pro 1. a 2. shluk. (Černě - 1. shluk, bíle - 2. shluk)

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 uvedeno v části „Vymezení problému“, 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.

Literatura

[1] M. S. Aldenderfer, R. K. Blashfield: Cluster Analysis. Sage Publications, Inc. London 1989.
[2] B. S. Everitt: Cluster Analysis. Edward Arnold, London 1998.
[3] J. C. Davis: Statistics and Data Analysis in Geology. John Wiley & Sons, Inc., 1973.
[4] P. Hebák, J. Hustopecký: Vícerozměrné statistické metody s aplikacemi. SNTL/Alfa, Praha 1987.
[5] J. Macek: Analýza znečištění ovzduší oxidem siřičitým a polétavým prachem ve městě Brně. Katedra geografie PřF MU Brno 1998.

MU PřF, KAM, Janáčkovo 2a, 662 95 Brno
E-mail: budikova@math.muni.cz