vodní hospodářství®

www.vodnihospodarstvi.cz

ročník 57
11
2007

1997-2007 EWA 10 let členství

20. - 21. 11. 2007 VODNÍ TOKY 2007,
5. ročník odborné konference s mezinárodní účastí,
hotel Černigov v Hradci Králové.
Pořádají Výbor ČVTVHS, VRV a.s. a Povodí Labe, státní podnik.
Další informace na www.pla.cz

PŘÍLOHA • ČL •

Statistické zpracování vodohospodářských dat

7. Přednosti analýzy shluků při klasifikaci zdrojů pitné vody

Milan Meloun

Klíčová slova
shluková analýza - dendrogram proměnných - dendrogram objektů - pitná voda - analýza vody - graf komponentního skóre - indexový graf vlastních čísel - graf komponentních vah - korelační matice

Souhrn

Analýza shluků (Cluster analysis) patří mezi metody, které se zabývají vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství znaků) a jejich klasifikací do tříd čili shluků. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. V biologii se proto užívá shluková analýza ke klasifikování živočichů a rostlin. Tato klasifikace se nazývá numerická taxonomie. Lze formulovat tři hlavní cíle analýzy shluků: popis systematiky, je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů, zjednodušení dat, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty, a konečně identifikace vztahu, kdy po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty.

◆

1. Úvod

Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková analýza neobsahuje techniku k rozlišení významných a nevýznamných znaků. Provede pouze odlišení shluků. Nesprávné zařazení znaků vede k zahrnutí i odlehlých objektů, které mohou mít rušivý vliv na výsledky analýzy. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

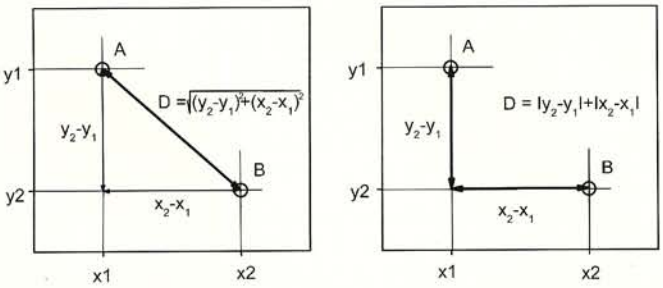
2. Vlastnosti metody

2.1 Identifikace odlehlých objektů

Při odhalování struktury objektů je shluková analýza velmi citlivá na přítomnost nevýznamných znaků. Je citlivá také na přítomnost odlehlých objektů, které se silně odlišují ode všech ostatních objektů. Odlehlé objekty mohou představovat buď (1) skutečně odchýlené, patologické objekty, které nejsou představiteli analyzované populace, nebo (2) chybný výběr objektu z populace, který způsobí nevhodné zastoupení původní populace.

2.2 Míry podobnosti

Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů. Nejdříve se stanovují znaky, určující podobnost, které se dále kombinují do podobnostních měr. Tímto způsobem pak může být objekt porovnán s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin:



Obr. 1. Nejpoužívanější míry vzdálenosti: (a) Eukleidovská *D*, (b) Manhattanovská vzdálenost *D*

(1) Korelační míry: Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále může být Pearsonův párový korelační koeficient r. Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedničce. V případě ordinální škály (pořadová čísla) je analogickou mírou podobností Spearmanův korelační koeficient. Obyčejně se vychází z transponované matice dat X^T, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice X^T představují korelaci mezi dvojicí objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou „podobnost" a nízká korelace pak „nepodobnost" profilů.

(2) Míry vzdálenosti **(obr. 1.)**: Představují nejčastěji užívané míry, založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Nejčastější vzdálenostní mírou je Eukleidovská vzdálenost zvaná také geometrická metrika, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že vzdálenost

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

představuje standardní typ vzdálenosti. Vedle Eukleidovské vzdálenosti se užívá také čtverec Eukleidovské vzdálenosti, který tvoří základ Wardovy metody shlukování. často je užívaná Manhattanská vzdálenost zvaná také vzdálenost městských bloků nebo Hammingova metrika, definovaná vztahem

$$d_H(x_k, x_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}|$$

Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. Další mírou je zobecněná Minkovského metrika, pro kterou platí

$$d_M(x_k, x_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z}$$

kde pro z = 1 jde o Hammingovu metriku a pro z = 2 o Eukleidovu. čím je z větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. V některých případech se používá také tětivová vzdálenost (anglicky chord distance), definovaná vztahem

$$d_{CH}(x_k, x_l) = \sqrt{2\left[1 - \frac{\sum_{j=1}^{m} x_{kj}\, x_{lj}}{\sum_{j=1}^{m} x^2_{kj} \sum_{j=1}^{m} x^2_{lj}}\right]}$$

V případě třech znaků je tětivová vzdálenost přímou vzdáleností dvou bodů na povrchu koule s jednotkovým poloměrem a počátkem v těžišti.

Problém všech vzdálenostních měr vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, díky často veliké odlišnosti jednotek měření. Shluky různých vzdálenostních měr se budou lišit, největší rozptýlení mezi shluky bude u čtverce Eukleidovské vzdálenosti. Pořadí podobností se významně změní se změnou měřítka nebo změnou jednotek jednoho ze znaků.

Všechny dosud uvedené metriky neuvažují závislost mezi znaky. Zahrneme-li do vztahu pro vzdálenost také vazby mezi znaky, vyjádřené kovarianční maticí C, dostaneme novou statistickou míru, zvanou Mahalanobisova metrika

$$d_{Ma}(x_k, x_l) = \sqrt{(x_k - x_l)^T C^{-1} (x_k - x_l)}$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální. Vysoce korelovaný výběr znaků může skrytě převážit celý soubor znaků shlukování.

(3) Míry asociace: Míry asociace podobnosti se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru (například binární proměnné). Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí ano nebo ne. Míra asociace pak vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti na danou otázku odpověděli ano nebo ne, tedy 1 nebo 0. Rozšíření tohoto jednoduchého „souhlasného koeficientu" je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků.

Uvedeme příklad: předpokládejme, že sledujeme asociaci mezi dvěma objekty O_i a O_j. Možné binární odezvy typu 0-1 je pak možno zapsat do tzv. kontingenční **tabulky 1.**:

Tab. 1.	Objekt O_i		
Objekt O_j		1	0
	1	a	b
	0	c	d

V této **tabulce 1.** jsou shrnuty všechny možné kombinace počtu znaků pro dva objekty:a značí počet znaků, kde mají oba objekty O_j a O_i hodnotu 1 a jde o tzv. pozitivní shodu,b značí počet znaků, kde má objekt O_j hodnotu 1 a objekt O_i hodnotu 0. c značí počet znaků, kde má objekt O_j hodnotu 0 a objekt O_i hodnotu 1. d značí počet znaků, kde mají oba objekty O_j a O_i hodnotu 0 a jde o tzv. negativní shodu. Míry asociace pak vyjadřují relativní podíly počtu znaků s ohledem na to, zda má smysl uvažovat negativní shodu nebo zda má nulová hodnota znaku u porovnávaných objektů stejnou příčinu. Mezi základní koeficienty podobnosti potom patří:

(a) Sokalův-Michenerův koeficient asociace (zvaný také koeficient jednoduché shody)

$$S_{SM} = \frac{a+d}{a+b+c+d}$$

(b) Russelův-Raoův koeficient asociace

$$S_{RR} = \frac{d}{a+b+c+d}$$

(c) Jaccardův koeficient se liší od předešlého Russelova-Raova koeficientu asociace jen tím, že má v čitateli počet shodných znaků a.

(d) Hamannův koeficient asociace

$$S_H = \frac{a+d-b-c}{a+b+c+d}$$

a také lze konstruovat následující obdobu.

(e) Korelační koeficient

$$r_B = \frac{a\,d - b\,c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(f) Rogersův a Tanimotův koeficient asociace je definován vztahem

$$S_{RT} = \frac{a+d}{a+2b+2c+d}$$

(g) Sörensenův koeficient asociace je definován vztahem

$$S_S = \frac{2a}{2a+b+c}$$

2.3 Standardizace dat

Před vlastní shlukovou analýzou je třeba řešit otázku, zda je třeba data standardizovat. Musí se respektovat fakt, že většina měr vzdálenosti je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají větší vliv na míru podobnosti.

(1) Standardizování znaků: Nejužívanější formou standardizace je normalizace každého znaku do svého Z-skóre, tj. odečtením průměru a dělením směrodatnou odchylkou. Tato standardizace je známa pod názvem normovací Z-funkce. Tato transformace eliminuje rozdíly v měřítku, mnohdy i řádově se lišících znaků. Výhody standardizace znaků jsou následující: (a) Znaky lze v jednotném měřítku, (kde je průměrná hodnota 0 a směrodatná odchylka 1) vzájemně porovnávat snadněji. Kladné hodnoty jsou nad průměrem a záporné hodnoty jsou pod průměrem, (b) Se změnou měřítka nedojde k rozdílu mezi standardizovanými znaky. I když změníme například u znaku čas jednotky z minut na sekundy, standardizované hodnoty budou stále stejné.

(2) Standardizace objektů: Když chceme identifikovat shluky dle vzdálenosti pak standardizace není vhodná. Standardizace objektů nebo-li řádková standardizace může být však efektní ve speciálních případech.

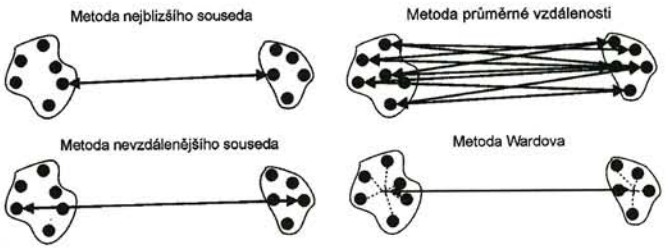
3. Způsoby shlukování

Shluk (cluster) je skupina objektů, jejichž vzdálenost je menší než vzdálenost s objekty do shluku nepatřícími. Podle způsobu shlukování se postupy dělí na hierarchické a nehierarchické. Hierarchické se dělí dále na aglomerační a divizní.

3.1 Hierarchické shlukovací postupy

Postupy jsou založeny na hierarchickém uspořádání objektů a jejich shluků. Graficky se hierarchicky uspořádané shluky zobrazují formou vývojového stromu nebo dendrogramu. U aglomeračního shlukování se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako objekt. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. Divizní postup je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů.

Mezi metody metriky shlukování patří – viz **obr. 2.**:



Obr. 2. Nejčastěji užívané metriky shlukování.

(1) Metoda nejbližšího souseda: Postup je postaven na minimální vzdálenosti. Naleznou se dva objekty, oddělené nejkratší vzdáleností a umístí se do shluku. Další shluk je vytvořen přidáním třetího nejbližšího objektu. Proces se opakuje až jsou všechny objekty v jednom společném shluku. Vzdálenost mezi dvěma shluky je definována jako nejkratší vzdálenost libovolného bodu v prvním shluku vůči libovolnému bodu v druhém. Dva shluky jsou propojeny v libovolném stádiu nejkratší spojkou.

(2) Metoda nejvzdálenějšího souseda: Kritérium je postaveno nikoliv na minimální ale na maximální vzdálenosti. Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích. Metoda se také nazývá metodou úplného propojení, protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti čili minimální podobnosti.

(3) Metoda průměrné vzdálenosti: Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

(4) Wardova metoda: Principem není optimalizace vzdáleností mezi shluky ale minimalizace heterogenity shluků podle kritéria minima přírůstku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku.

(5) Metoda těžiště: Jde o vzdálenost dvou těžišť shluků, vyjádřených Eukleidovskou vzdáleností nebo čtvercem Eukleidovské vzdálenosti. Těžiště shluku má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé znaky. Po každém kroku shlukování se počítá nové těžiště. Poloha těžiště shluku poněkud migruje tak jak se připojují nové objekty a vznikají větší shluky.

(6) Metoda mediánová: Jde o jisté vylepšení metody těžiště, neboť se snaží odstranit rozdílné významnosti, které metoda těžiště dává různě velkým shlukům.

3.2 Nehierarchické shlukovací postupy

U metody zárodečných bodů (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají tvořit zárodky nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich Eukleidovské vzdálenosti od těchto typických objektů. Existuje několik postupů zadávání zárodků shluku a zařazování objektů do shluku. Těmto metodám se říká K-means shlukování:

(a) Sekvenční práh: Metoda začíná volbou jednoho zárodku shluku a zahrnuje všechny objekty uvnitř předspecifikované vzdálenosti. Když jsou všechny objekty uvnitř této vzdálenosti zahrnuty do shluku, je vybrán zárodek druhého shluku a všechny objekty uvnitř předspecifikované vzdálenosti jsou zahrnuty do shluku. Pak je vybrán třetí zárodek shluku a proces se opakuje. Když je jednou objekt shlukován se zárodkem, není s ním více počítáno do některého jiného shluku.

(b) Paralelní práh: Na rozdíl od předešlého postupu tento postup vybírá na začátku několik shlukových zárodků současně čili paralelně a zařazuje objekty uvnitř prahové vzdálenosti do nejbližšího zárodku. Jak se proces vyvíjí, prahovou vzdálenost lze nastavit tak, aby zařadila více nebo méně objektů do shluku. Některé objekty mohou zůstat nezařazeny do shluků, když se totiž nacházejí vně předspecifikované vzdálenosti od shlukového zárodku.

(c) Optimalizace: Metoda zvaná optimalizační postup je podobná předešlým dvěma kromě toho, že dovoluje znovuzařazení objektů. Když se v průběhu tvorby shluků stane, že některý objekt se octne blíže jinému shluku, než ve kterém se právě nachází, optimalizační postup ho přeřadí do jiného bližšího shluku.

3.3 Dendrogramy hierarchického shlukování

Analýzou shluků je možné hodnotit jednak podobnost objektů, analyzovanou pomocí dendrogramu objektů, a jednak podobnost znaků analyzovanou pomocí dendrogramu znaků.

Dendrogram shluků (vývojový strom) se konstruuje pouze v případě, kdy je k dispozici matice původních znaků. Dendrogram podobnosti znaků ukazuje rozlišení znaků ve shlucích. Jeho interpretace je snadná: znaky blízko sebe jsou propojeny spojovací úsečkou hodně nízko, mají malou vzdálenost čili značnou vzájemnou podobnost. Znaky propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (obecně m-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku si jsou značně podobné a jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: Dendrogram lze sestrojit celou řadou technik. Prvním kritériem těsnosti proložení při volbě „nejlepšího dendrogramu", jež nejlépe odpovídá struktuře objektů a znaků mezi objekty, je kofenetický korelační koeficient CC. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

Druhým kritériem těsnosti proložení je kritérium delta Δ, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d^*_{jk}|^{1/A}}{\sum_{j<k}^{N} (d^*_{jk})^{1/A}} \right]^A$$

kde $A = 0.5$ nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu. Je žádoucí, aby hodnoty *delta* byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

4. Shlukování metodou nejbližších těžišť (K-Means)

Metoda nejbližších těžišť poskytuje pouze jediné řešení pro počet požadovaných shluků. Počet shluků musí být zadán uživatelem. Postup je založen na nejbližším těžišti, kdy objekt je zařazen do shluku s nejmenší vzdáleností mezi objektem a těžištěm shluku. Jinak jsou těžiště shluků určována iteračně z dat.

1. Klasifikace, když těžiště shluků jsou známa: Metoda požaduje uživatelem zadaný počet shluků, proto je třeba nejprve aplikovat hierarchickou analýzu shluků na náhodný výběr objektů a určit počet shluků.

2. Klasifikace, když těžiště shluků nejsou známa: Existuje celá řada případů, kdy těžiště shluků nejsou předem známa. Správná těžiště dobře separují shluky. V následujících krocích nahradí objekt těžiště, když jeho nejmenší vzdálenost k těžišti bude větší než vzdálenost mezi dvěma nejbližšími těžišti.

4.1 Shlukování metodou optimálních medoidů

Medoid, čili optimální střed shluku, je střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Je-li požadováno k shluků, bude existovat také k medoidů. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu. Medoidy a shluky se vytvářejí na základě vzdáleností čili nepodobností.

Druhů možných znaků v tomto případě shlukování je celá řada: Intervalové jsou spojité kladné či záporné, v lineární škále, např. výška, hmotnost, cena, teplota, čas atd. Ordinální jsou pořadová čísla stupnice, hodnotící nějakou vlastnost, např. silný nesouhlas (5), nesouhlas (4), neutrální (3), souhlas (2) a silný souhlas (1). Poměrové jsou kladné hodnoty, kdy vzdálenost mezi dvěma čísly je stejná, když i jejich poměr je stejný, např. mezi 3 a 30 je stejná jako mezi 30 a 300, chemická koncentrace, intenzita záření, absorbance atd. Nominální jsou znaky, které vyjadřují pouze kvalitu a nikoliv kvantitu, např. PSč, rasa, barva, název města atd. Symetrické binární: mají dvě možnosti, obvykle ano (1), ne (0). Asymetrické binární: přítomnost či nepřítomnost zřídka se vyskytující události, kdy nepřítomnost není tak důležitá, např. osoba má jizvu na tváři, a tím je lépe identifikovatelná.

4.2 Späthova metoda

Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná u počátečního uspořádání shluků, algoritmus pak najde lokální minimum inteligentním přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, proces končí. Lokální minimum však nemusí být globálním. Aby program překonal toto omezení, zopakuje se několikrát hledání vždy z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Jako účelová funkce se bere celková vzdálenost mezi všemi objekty ve shluku podle vzorce

$$D = \sum_{l=1}^{k} \sum_{i \in c_k} \sum_{j \in c_k} d_{ij}$$

kde k je celkový počet shluků, d_{ij} je vzdálenost mezi i-tým a j-tým objektem a c_k je soubor objektů ve shluku l.

4.3 Silueta

Silueta je statistika, která poskytuje klíčovou informaci o dobrém a špatném shluku. Hodnota siluety s se vypočte tímto postupem:

1. Objekt i je ve shluku A a má průměrnou vzdálenost a ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, je a = 0.

2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu i ve shluku A a b je průměrná vzdálenost mezi objektem i a všemi objekty ve shluku B.

3. Silueta s objektu i se vyčíslí: když shluk A obsahuje pouze jeden objekt, je s = 0. Když a < b, je s = 1 - a/b. Když a> b, je s = b/a - 1. Když a = b, je s = 0.

Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů nejbližšího souseda B dle pravidla:

1. Je-li s blízko hodnotě +1, objekt i je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda B.

2. Je-li s blízko hodnoty nula, objekt i se nachází kdesi uprostřed mezi shluky A a B, a čistě náhodou byl přiřazen do shluku A.

3. Je-li s blízko hodnotě -1, objekt i je špatně klasifikován. Vzdálenosti k ostatním objektů ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda B. Otázkou pak je, proč byl vlastně zařazen do shluku A.

4.4 Určení počtu shluků

Přehlednou statistikou je průměrná silueta s, počítaná přes všechny objekty. Tato hodnota sumarizuje jak těsně prokládá shlukové uspořádání analyzovanými daty. Nalezení správného počtu shluků spočívá v nalezení takového počtu, který maximalizuje průměrnou siluetu. Označme maximální hodnotu průměrné siluety všech shluků k symbolem SC a pak budeme rozlišovat následující typy shlukových uspořádání – **Tab. 2.**:

Tabulka 2.

SC	Vysvětlení uspořádání do shluků
od 0.71 do 1.00	Silná a dobrá struktura.
od 0.51 do 0.70	Ještě přijatelná struktura.
od 0.26 do 0.50	Slabá struktura, asi umělá. Je třeba najít novou, lepší.
od -1.00 do 0.25	Naprosto nevhodná struktura.

5. Postup obecné analýzy shluků

Analýza shluků je vždy silnou analytickou pomůckou k účelům zjednodušení, průzkumu a potvrzení struktury.

1. krok: Cíle analýzy shluků: Primárním cílem je rozdělení souboru objektů do dvou nebo více skupin, tříd či shluků. Sledujeme tři cíle:

(a) Popis systematiky: Tradičním využitím jsou průzkumové cíle a popis systematiky - taxonomie, tj. empirická klasifikace objektů. Analýzou shluků se dospěje k určitým shlukům objektů, které jsou pak porovnány s jejich teoreticky odvozenou typologií.

(b) Zjednodušení dat: Při hledání taxonomie poskytuje analýza shluků zjednodušený pohled na objekty. Zatímco faktorová analýza se snaží nalézt strukturu znaků, analýza shluků činí totéž ale pro objekty. Na objekty se pak už nehledí jako na jeden společný soubor ale na oddělené shluky objektů, rozčleněné dle jejich vlastností.

(c) Identifikace vztahu: Po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty, což by bylo mezi samotnými objekty daleko obtížnější. Shluky mohou být předmětem dalšího kvalitativního uvažování.

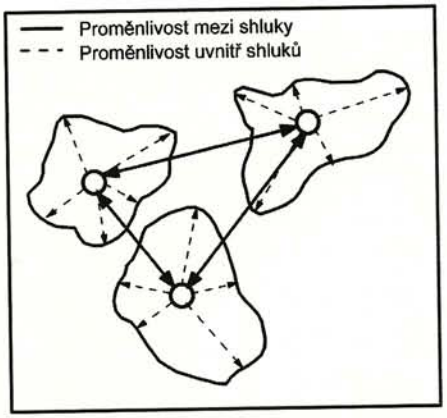
2. krok: Formulace úlohy analýzy shluků: S vybranými znaky budeme shlukovat vyšetřované objekty. Nejprve však je třeba odpovědět na tři otázky: (1) Mohou být v datech nějaké odlehlé objekty, které mohou být posléze odstraněny? (2) Jak vyjádříme podobnost objektů? (3) Měla by být data před analýzou shluků standardizována?

3. krok: Předpoklady analýzy shluků – obr. 3: Analýza shluků objektů není charakteru statistického testování. Požadavky normality, linearity, homoskedasticity, které jsou tolik důležité v ostatních vícerozměrných

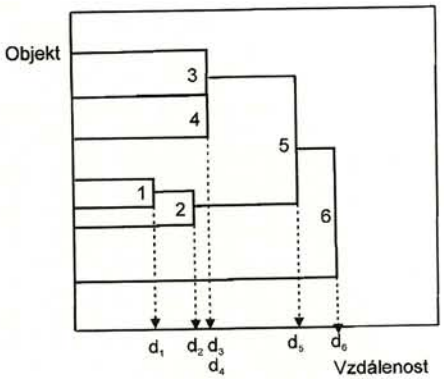
technikách nemají zde význam. Přesto existují dva kritické předpoklady: reprezentativnost a vliv multikolinearity.

Reprezentativnost vzorku: Předpokládá se, že výběr objektů a odvozené shluky reprezentují strukturu celého souboru. Uživatel si proto musí být jist, že zvolený výběr dat je opravdovým představitelem sledovaného souboru.

Vliv multikolinearity: Multikolinearita se chová jako neviditelný proces vážení, který silně ovlivňuje analýzu. Uživatel musí proto ověřit přítomnost multikolinearity a když je tato prokázána, je třeba zredukovat počet znaků nebo použít vhodnou míru, jako např. Mahalanobisova vzdálenost, která je schopná multikolinearitu kompenzovat.



Obr. 3. Porovnání vzdálenosti mezi shluky a uvnitř shluku.



Obr. 4. Postupná výstavba dendrogramu.

4. krok: Výstavba shluků a celková těsnost proložení: Za vhodné rozlišovací kritérium je možné použít maximalizaci rozdílů mezi shluky, a toto porovnávat vůči proměnlivosti uvnitř shluků. Testování poměru roztylu mezi jednotlivými shluky vůči průměru rozptylu uvnitř shluku lze porovnat s F kritériem v analýze rozptylu.

Hierarchické shlukování: Tyto algoritmy se týkají konstrukce stromovité struktury shluků, dendrogramu – **obr. 4.** Existují v zásadě dva typy hierarchického shlukování, aglomerační a divizní. Grafickým zobrazením růstového stromu je diagram, také zvaný dendrogram. Jinou grafickou metodou je vertikální rampouchovitý diagram – **obr. 5**.

Když shlukovací proces probíhá v opačném směru než aglomerační, označuje se jako metoda divizní. Postup začíná z jednoho velkého shluku, ve kterém jsou všechny objekty. V následujících krocích jsou nepodobné objekty odloučeny ze společného shluku a vzniká tím menší shluk. Proces probíhá tak dlouho, až je ve shluku jediný objekt. Pět nejpoužívanějších aglomeračních algoritmů výstavby shluků jsou metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, metoda průměrová, Wardova metoda, a metoda těžiště.

Počet vytvářených shluků: Snad nejvíce matoucí otázkou v analýze shluků je dosažení konečného počtu shluků, známého také pod názvem terminační kritérium. Neexistuje žádný objektivní způsob určení tohoto kritéria. Jedno z terminačních kritérií se týká relativně jednoduchého vyšetření měr podobnosti mezi shluky v každém kroku, když totiž míra podobnosti překročí předdefinovanou velikost nebo když následné hodnoty se skokově změní. Je vhodné postupovat tak, že se určí rozličný počet shluků např. 2, 3 a 4 a na základě úvah o alternativním řešení, praktickém úsudku a teoretických základech úlohy samé se rozhodne. Když se objeví jednoobjektový shluk nebo shluk o poměrně malé velikosti, uživatel musí rozhodnout, zda tento představuje strukturního člena vzorku nebo zda ho lze označit jako nedostatečně reprezentativní pro soubor dat.

5. krok: Interpretace shluků: Interpretace shluků se týká vyšetření každého shluku v pojmech shlukových znaků a především pojmenování shluků, které vystihuje podstatu a povahu shluků. Při zahájení interpretace si uvědomíme, že ve shlukové analýze je často užívanou mírou těžiště shluků. Uživatel se musí vrátit k původním datům v původních znacích a vyčíslit průměrové profily pro původní data. Vyšetříme proto profily průměrových skóre a označíme popisným nadpisem každý shluk.

6. krok: Validace a profilování shluků: Existuje poněkud subjektivní charakter analýzy shluků stran hledání optimálního shlukového řešení.

Validace shluků: Validace čili potvrzení shluků zahrnuje pokusy zajistit, že shlukové řešení je reprezentativní v celé obecné populaci, a je proto zobecnitelné i na ostatní objekty a dále je i stabilní v čase.

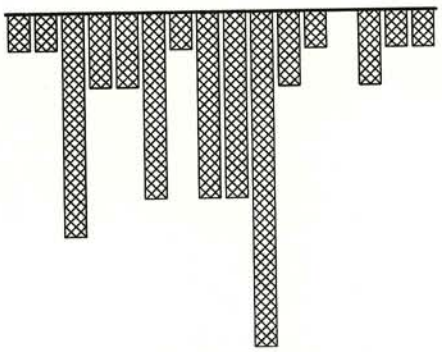
Profilování shlukového řešení: Profilování se týká popisu vlastností každého shluku, aby se objasnilo jak se vlastnosti liší ve významných dimenzích. Profilová analýza se zaměřuje na popis ne toho, co konkrétně odhalují shluky ale na vlastnosti shluků, na jejímž základě byly identifikovány. Kromě toho je kladen důraz na vlastnosti, ve kterých se shluky vzájemně odlišují a na vlastnosti, které mohou predikovat účast v dotyčném shluku.

Vzorová úloha 3. Klasifikace zdrojů pitné vody (E404)

Na 62 vzorcích zdrojů pitné vody bylo experimentálně stanoveno 16 znaků čili proměnných kvality vody. Je třeba vyšetřit, zda lze nalézt vybočující objekty, resp. jejich znaky, zda existuje korelace mezi znaky, zda ukazuje graf komponentních vah na korelující znaky, zda lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, zda lze posoudit podobnost objektů shlukovou analýzou klasifikaci zdrojů.

Data: i index vzorku, sledované znaky jsou **NO3** obsah dusičnanů [mg/l], **NO2** obsah dusitanů [mg/l], **Cl** obsah chloridů [mg/l], **Cl2** obsah celkového chloru [mg/l], **SO4** obsah síranů [mg/l], **PO4** obsah fosforečnanů [mg/l], **NH4** obsah amonných solí [mg/l], **Ca** obsah vápníku [mg/l], **Mg** obsah hořčíku [mg/l], **Fe** obsah železa (celkového) [mg/l], **Mn** obsah manganu [mg/l], **pH** je pH roztoku, **KNK**, **ZNK**, **Vodiv.** vodivost roztoku, **Nerozp.** nerozpuštěné látky [mg/l].

(Tabulka zdrojové matice dat je uvedena v předešlém sdělení této řady a také k dispozici na internetu na adrese: //meloun.upce.cz/data a dále pak v bloku Kompendium-Data).



Obr. 5. Výstavba rampouchovitého diagramu.

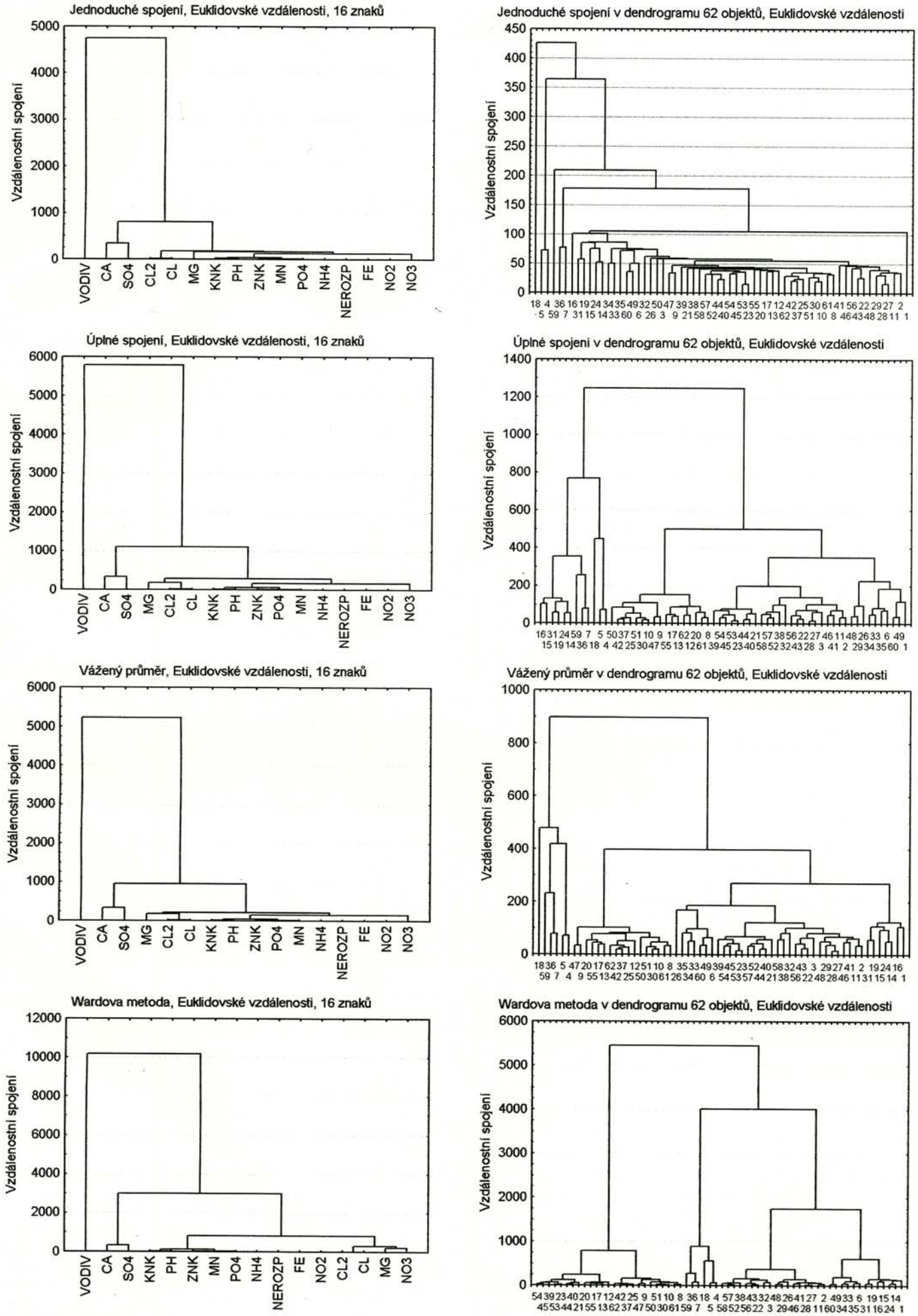
Řešení:

V rámci úvodní průzkumové analýzy zdrojové matice dat je nejprve exploratorně aplikována metoda hlavních komponent. Cattelův indexový graf vlastních čísel korelační matice ukázal zlom u 2 a 3 komponent. Z toho důvodu budou nadále využívány grafické diagnostiky od prvních tří komponent a porovnávány s dendrogramem shlukové analýzy – **obr. 6. a 7.**

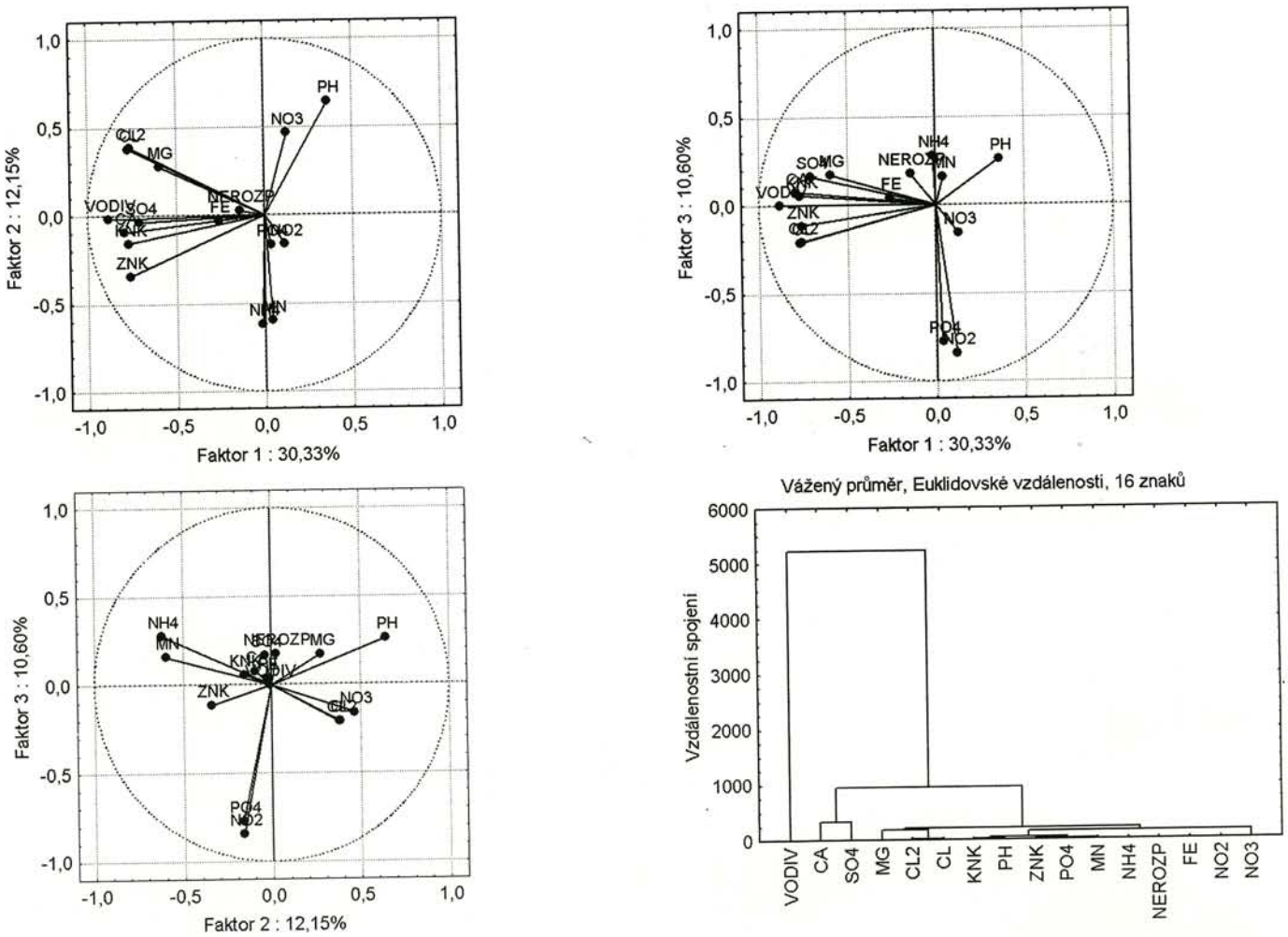
Protože cílem této úlohy je prezentace shlukové analýzy, bude dendrogram proměnných (v levé části **obr. 6.**) postupně vytvořen čtyřmi rozličnými technikami. Dendrogram zde ukazuje na 4 shluky a jednu vybočující odlehlou proměnnou - vodivost, která je značně nepodobná ostatním znakům čili obsahům iontů a molekul v pitné vodě. Největší shluk obsahuje znaky PO4-NH4-NEROZP-FE-NO2 a k němu se pak blíží i druhý shluk s obsahem KNK-PH-ZNK-MN. Oba shluky pak s třetím shlukem CL2-CL-MG-NO3 vytváří jeden velký společný shluk, který je zcela nepodobný shluku obsahujícímu CA-SO4. Dosud uvedené shluky lze spojit v jeden společný, vůči kterému je silně nepodobný znaku VODIV, stojící zde v roli odlehlé vybočující proměnné.

Dendrogram proměnných metodou vážených průměrů Eukleidovských vzdáleností 16 znaků se jeví dle rozhodčího kritéria kofenetického korelačního korelačního koeficientu jako nejlepší a bude proto prováván s grafy komponentních vah 1.-2., 1.-3. a 2.-3. (**obr. 7.**) hlavní komponenty. Graf 1. a 2. hlavní komponenty na **obr. 7a.** ukazuje, že silně spolu korelují dvojice či shluky proměnných MN-NH4, CL-CL2-MG, CA-SO4-KNK-ZNK, FE-NEROZP, PO4-NH4, NO3-PH. Naprosto nekorelují především shluk MN-NH4 se shlukem CL-CL2-MG, nekoreluje MN-NH4 se shlukem NO3-PH, nekoreluje PH-NO3 se shlukem CA-SO4-KNK-ZNK. Podobně lze určit i shluky v grafech **obr. 7b** a **7c**. Nejvíce proměnlivosti popisuje 1. hlavní komponenta (a to 30%), potom 2. hlavní komponenta (12%) a konečně nejméně 3. hlavní komponenta (10%). Zbytek proměnlivosti je zde obsažen ve zbývajících ostatních 4. a 16. komponentách, které jsou však spíše šumového charakteru.

Podobnou technikou vůči metodě hlavních komponent je faktorová analýza. Zde je užitečná jistá možnost této metody, a to rotace os latentních proměnných (faktorů) a získání tak zobrazení faktorově čistých proměnných. Na **obr. 8.** jsou tři grafy faktorových vah pro 1.-2., 1.-3. a 2.-3. faktory po Varimax rotaci faktorových os. Umístění proměnných na ploše je pak přehlednější, protože rotace zajišťuje faktorově čisté čili jednou osou dostatečně oddělené znaky. V grafech lze detekovat následující shluky:NH4-MN, PO4-FE-NO2-NEROZP, MG-KNK-ZNK, CA-SO4-VODIV a CL-CL2. Zcela osamoceně se nachází PH a NO3, které mají v analýze piné vody poněkud výjimečné postavení. Uvedené shluky jsou přehledně zobrazeny také na 3D grafu na **obr. 8d.**



Obr. 6. Vlevo dendrogram proměnných, vpravo dendrogram objektů. Metody shora dolů: nejbližšího souseda, nejvzdálenějšího souseda, váženého průměru a Wardova, STATISTICA.



Obr. 7. Grafy komponentních vah pro (a) 1.-2. komponenty, (b) 1.-3. komponenty, (c) 2.-3. komponenty, a (d) dendrogram proměnných metodou váženého průměru pro 16 znaků, STATISTICA.

Obr. 9. přináší rozptylový diagram komponentního skóre pro 1.-2., 1.-3. a 2.-3. komponenty. Přehled rozličných matematických přístupů je rovněž v pravé části diagramu na **obr. 6.** Komponentní skóre je zde porovnáváno se shlukem objektů metodou váženého průměru v dendrogramu 62 objektů čili vzorků pitné vody. Vedle jednoho velkého společného shluku jsou v dendrogramu detekovány odlehlé body vzorků 18, 36, 59, 7, 5, 4, což je převážně ve shodě s detekovanými odlehlými body komponentním skórem. Je třeba si uvědomit, že jde o matematicky zcela jinou nezávislou metodu než je metoda výstavby dendrogramu na základě podobností.

Závěr

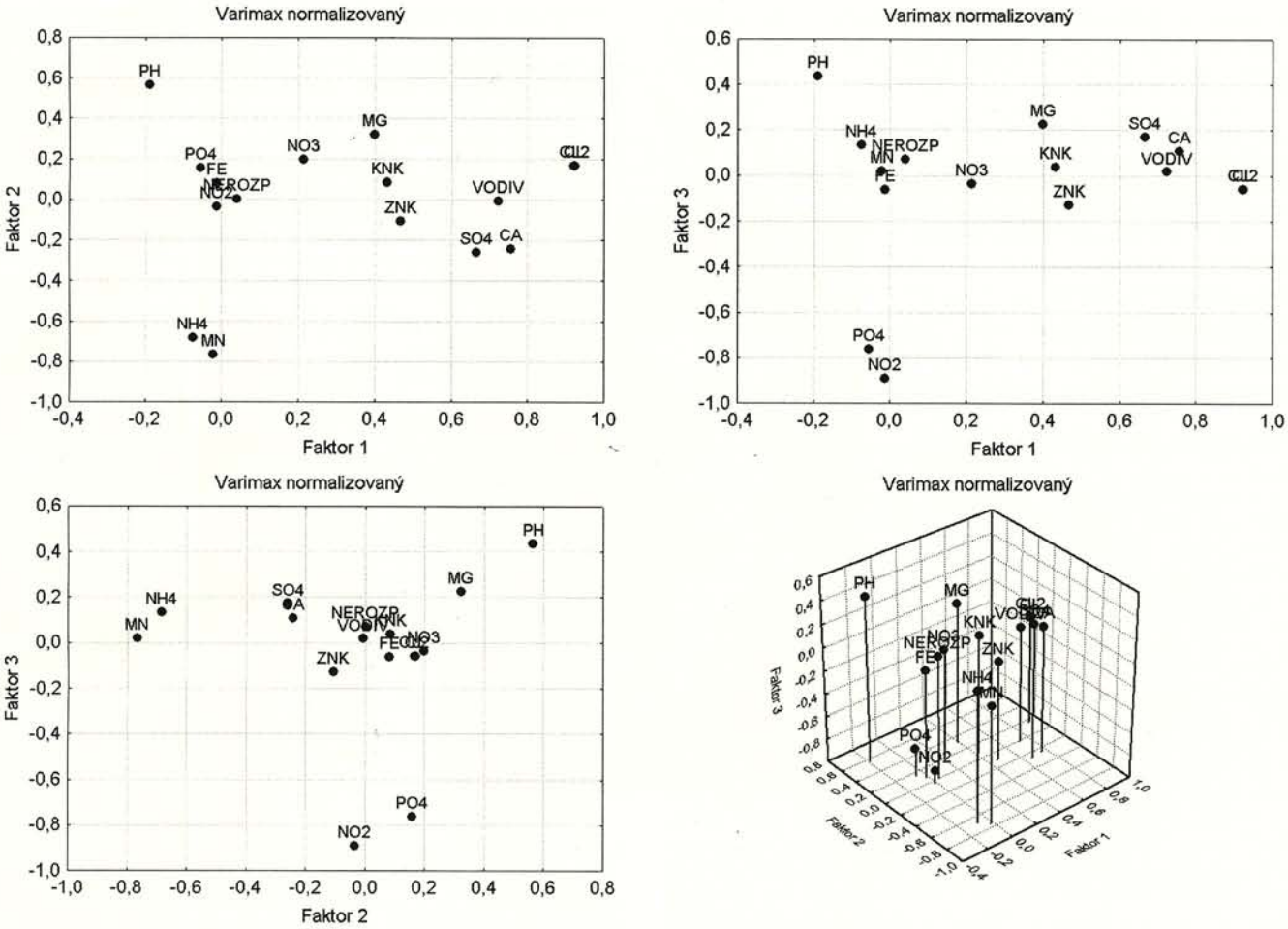
Shlukováním lze v dendrogramu znaků provést klasifikaci naměřených a sledovaných znaků pitné vody a v dendrogramu objektů pak klasifikaci vzorků pitné vody.

Poděkování: Autoři vyslovují svůj dík za finační podporu vědeckého záměru č. MSM0021627502.

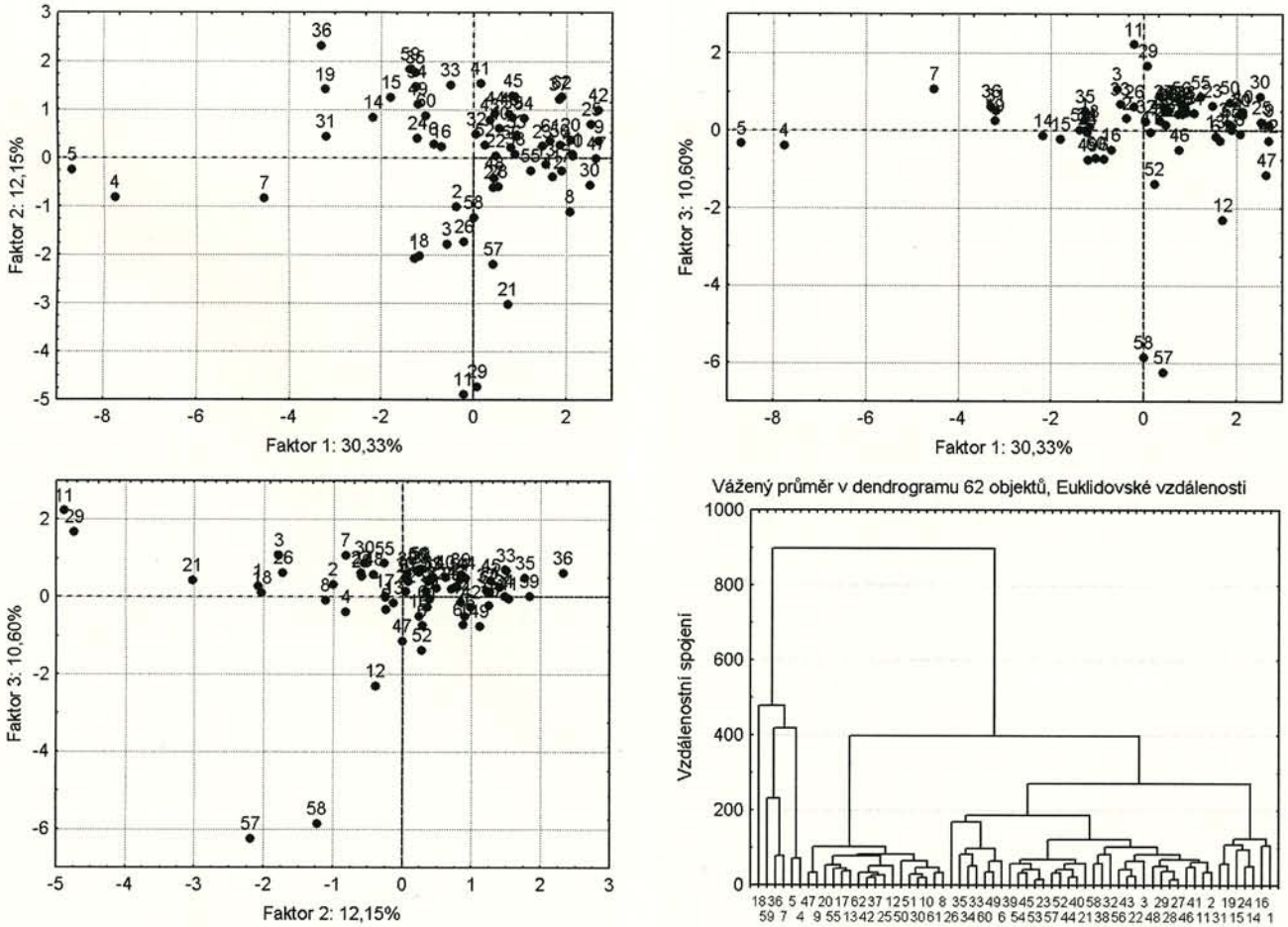
Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: Modern Multivariate Statistical Analysis, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Kendall M. G., Stuart A.: The Advanced Theory of Statistics, Vol. III. New York 1966.
[3] James W., Stein C.: Estimation with Quadratic Loss, Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.
[4] Guanadeskian R., Kettenring J. R.: Biometrics **28**, 80 (1972).
[5] Campbell N. A.: Appl. Statist., 29, 231 (1980).
[6] Hu J., Skrabal P., Zollinger H.: Dyes and Pigments, **8**, 189 (1987).
[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: Graphical Methods for Data Analysis. Duxburg Press, Belmont, California 1983.
[8] Barnett V., (Edit.): Interpreting Multivariate Data. Wiley, Chichester 1981, kap. 6.
[9] Jolliffe I. T.: Principal Component Analysis. Springer Verlag, New York 1986.
[10] Barnett V., (Edit.): Interpreting Multivariate Data. Wiley, Chichester 1981, kap. 12.
[11] Everitt B. S.: Graphical Techniques for Multivariate Data. London 1978.
[12] Andrews D. F.: Biometrics, **28**, 125 (1972).
[13] Kulkarni S. R., Paranjape S. R.: Commun. Statist., **13**, 2511 (1984).
[14] Guanadeskian R.: *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York 1977.
[15] Kleiner B., Hartigan J. A., *J. Amer. Statist. Assoc.*, **76**, 260 (1981).
[16] Kres H.: *Statistical Tables for Multivariate Analysis.* Springer, New York 1983.
[17] Seber G. A. F.: *Multivariate Observations.* Wiley, New York 1984.
[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: *Z. Anal. Chem.*, **327**, 679 (1987).
[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: *Technometrics*, **24**, 139 (1982).
[20] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982
[21] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy*, SNTL Praha 1981
[22] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester 1992.
[23] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[24] Krzanowski W. J.: *Principles of Multivariate Analysis, A User=s Perspective*, Oxford Science Publications 1988,
[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).
[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004.
[27] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.
[28] Thomas E. V., *Anal. Chem.*, **66** (1994) 795A-804A.
[29] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.
[30] Meloun M. , Militký J., Kompendium statistického zpracování experimentálních dat, Academia Praha 2002, Academia Praha 2006.

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Chemickotechnologická fakulta,
Univerzita Pardubice,
nám. Čs. Legií 565, 532 10 Pardubice,
http: //meloun.upce.cz,
email: milan.meloun@upce.cz,
telefon: 466 037 026, fax: 466 037 068,
ICQ: 224-001-003



Obr. 8. Grafy faktorových vah po Varimax rotaci pro (a) 1.-2. faktory, (b) 1.-3. faktory, (c) 2.-3. faktory, a (d) 3D graf prvních tří faktorů, STATISTICA.



Obr. 9. Grafy komponentního skóre po Varimex rotaci pro (a) 1.-2. komponenty, (b) 1.-3. komponenty, (c) 2.-3. komponenty, a (d) dendrogram objektů metodou váženého průměru pro 62 objektů, STATISTICA.

Computer-Assisted Statistical Data Analysis. 7. Advantages of Cluster Analysis in the Classification of Drinkable Water (Meloun, M.)

Key Words
cluster analysis - CLU - dendrogram of variables - dendrogram of objects - drinkable water - water analysis - potable water - scatterplot - scree plot - factor analysis - principal components analysis - components weight plot

The cluster analysis leads to clusters which may be plotted in dendrogram. There are two dendrograms available, the dendrogram of variables and the dendrogram of objects. Both statistical techniques are demonstrated on the analysis and classification of various sources of a drinkable water. Data matrix contains objects in n rows and m columns. Before data treatment the data are scaled. Similarity of objects and variables is considered on base on Mahalonobis distance or Euclidean distance in the m-dimensional space. The principal components analysis reduces dimensionality and presents objects in two or three dimensions. The plot of components weight shows hidden structure among variables while the scatterplot shows the hidden structure of objects.

Mezinárodní Symposium firmy Hans Huber AG WATER SUPPLY AND SANITATION FOR ALL Berching, SRN, 27. – 28. září 2007

Symposium uspořádala firma Huber ve spolupráci s panelem předních odborníků z celého světa. Toto symposium navázalo na dvě předchozí akce s celosvětovým ohlasem, která se pod patronací firmy Huber konala v Berchingu: 2002 – symposium o čištění odpadních vod; 2004 – symposium DeSaR o decentralizovaných systémech sanitace. Při příležitosti tohoto symposia byl slavnostně uveden do provozu systém DeSaR na zpracování odpadních vod ze správní budovy firmy. Tento systém je v provozu dodnes a produkty jsou mj.používány na hnojení stromů a rostlin ve firemním sadu a zahradě.

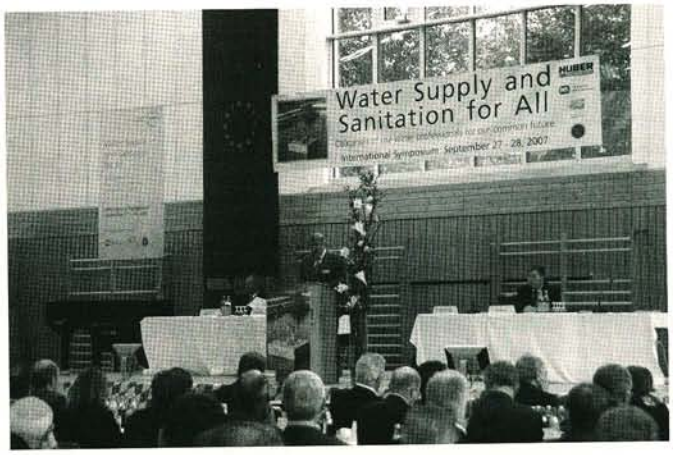
Program symposia byl směřován k nalezení odpovědi na otázku: **Dochází ke změně paradigmat?** K takové zásadní otázce vede rozbor faktorů, které dnes ovlivňují hospodaření s vodou v celosvětovém měřítku, jakými jsou zejména změny klimatu, dynamický nárůst populace, obrovský rozsah migrace, expanze měst do nebývalých rozměrů, atd. Za těchto nových okolností je zapotřebí si zodpovědět i další otázky, jako např. je rozumné budovat stejnou infrastrukturu pro zásobování pitnou vodou i sanitaci v regionech s různým klimatem, s různými ekologickými i ekonomickými podmínkami? Je skutečně trvale udržitelný náš splachovací koncept sanitace v době, kdy se čerstvá voda stává limitujícím zdrojem rozvoje? Neexistují snad chytřejší a rozumnější metody nakládání s našimi drahocennými vodními zdroji? A jsou vůbec vodohospodářské orgány, městští plánovači, architekti, regulátoři i politikové připraveni přijímat taková inovativní řešení, která se mnohdy odchylují od zažitých pouček v příručkách.

K řešení výše uvedených otázek se sjelo do Berchingu 500 účastníků, přičemž polovina byli delegáti z nejrůznějších zemí všech kontinentů, zastupující jak rozvinuté tak rozvíjející se ekonomiky. Přednášky byly rozděleny do několika tematických sekcí. Každou sekci vedl předseda delegovaný programovým výborem. Jeho úkolem bylo kromě moderování sekce uvést posluchače do probírané problematiky a shrnout výsledky diskuse. Náměty jednotlivých sekcí byly následující:

- Vytváření rámce
- Čištění pro opětovné používání
- Od centrálního k decentralizovanému
- Vypouštění pro opětovné upotřebení
- Řešení středních podniků – inovativní a trvale udržitelná řešení

Zájemci o jednotlivé přednášky prezentované na symposiu mohou nalézt většinu prezentací v PDF formátu na webové stránce:
http://www.huber.de/hp126032/Download-presentations-_International-Symposium_.htm

J. W.



§ PRÁVNÍ § PŘEDPISY

267/2007 Vyhláška, kterou se mění vyhláška Ministerstva životního prostředí České republiky č. 395/1992 Sb., kterou se provádějí některá ustanovení zákona České národní rady č. 114/1992 Sb., o ochraně přírody a krajiny, ve znění pozdějších předpisů, a výnos Ministerstva kultury České socialistické republiky č. j. 14.200/88-SÚOP ze dne 29. listopadu 1988 (reg. v částce 49/1988 Sb.)

262/2007 Nařízení vlády o vyhlášení závazné části Plánu hlavních povodí České republiky

219/2007 Nařízení vlády, kterým se mění nařízení vlády č. 103/2003 Sb., o stanovení zranitelných oblastí a o používání a skladování hnojiv a statkových hnojiv, střídání plodin a provádění protierozních opatření v těchto oblastech

216/2007 Zákon, kterým se mění zákon č. 100/2001 Sb., o posuzování vlivů na životní prostředí a o změně některých souvisejících zákonů (zákon o posuzování vlivů na životní prostředí), ve znění pozdějších předpisů

209/2007 Vyhláška, kterou se mění vyhláška Ministerstva dopravy a spojů č. 241/2002 Sb., o stanovení vodních nádrží a vodních toků, na kterých je zakázána plavba plavidel se spalovacími motory, a o rozsahu a užívání povrchových vod k plavbě, ve znění vyhlášky č. 39/2006 Sb.

180/2007 Zákon, kterým se mění zákon č. 86/2002 Sb., o ochraně ovzduší a o změně některých dalších zákonů (zákon o ochraně ovzduší), ve znění pozdějších předpisů

146/2007 Nařízení vlády o emisních limitech a dalších podmínkách provozování spalovacích stacionárních zdrojů znečišťování ovzduší

122/2007 Vyhláška, kterou se mění vyhláška č. 545/2002 Sb., o postupu při provádění pozemkových úprav a náležitostech návrhu pozemkových úprav