Odhalení skryté struktury a vnitřních vazeb dat vícerozměrnou statistickou analýzou pitné vody

Prof. RNDr. Milan Meloun, DrSc,

Katedra analytické chemie,
Univerzita Pardubice, 532 10 Pardubice,
milan. meloun@upce. cz

Souhrn: *Na výběru pitné vody jsou ukázány diagnostické vlastnosti vybraných metod vícerozměrné statistické analýzy, především metody hlavních komponent, faktorové analýzy a tvorby shluků. Uvedenými metodami je odhalena vnitřní struktura v datech, skryté vnitřní vazby a řada souvislostí ukrytých mezi znaky a mezi objekty.*

Odhalení struktury ve znacích a objektech: Zdrojová matice dat má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *průzkumovou (exploratorní) analýzu dat,* která umožňuje: a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafu, b) nalézt *vybočující objekty,* resp. jejich znaky, c) stanovit, zda lze použít předpoklad *lineárních vazeb,* d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).
Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda hledají 1. *strukturu a vazby ve znacích* nebo *2. strukturu a vazby v objektech.* To znamená a) hledání struktury ve znacích v metrické škále *-faktorová analýza FA, analýza hlavních komponent PCA* a *shluková analýza,* b) hledání struktury v objektech v metrické škále *- shluková analýza.*
Určením struktury a vzájemných vazeb mezi znaky ale i mezi objekty se zabývají techniky redukce znaků na latentní proměnné, *metoda* analýzy hlavních komponent *(PCA), metoda faktorové analýzy (FA)* a shlukování *(CLU).*

Maticový diagram. Diagram ukazuje rozptylové diagramy jednotlivých dvojic znaků. Je zřejmé, že vysoké hodnoty korelačního koeficientu vedou ke zřetelné lineární závislosti (přímka), zatímco nízké hodnoty ukazují, že osoby nejsou v grafu zobrazeny na přímce, ale nacházejí se spíše v chaotickém mraku bodů.

Směry maximálního rozptylu v rozptylovém diagramu. Intuitivně tušíme, že by bylo nejlepší proložit objekty novou souřadnou osu tak, aby byla totožná s nejtěsněji prokládanou přímkou. Nejtěsněji objekty prokládaná přímka, a tím pádem i nová souřadnicová osa zvaná *PCX,* leží ve směru *největšího rozptylu objektů.* Souřadnicová osa *PCI* není obecně přitom paralelní s žádnou z původních os znaků x_1, x_2 a x_3. Novou souřadnou osu *PCI* nazveme v následujících kapitolách *první hlavní komponentou.* Bude ležet *ve směru maximálního rozptylu* zdrojové matice dat **X.** S touto osou je spjata nová latentní proměnná, jejíž význam zatím neznáme. Metoda hlavních komponent

PCA nám poskytne nejenom tuto první komponentu, ale také další hlavní komponenty a je jenom na nás, jak je budeme interpretovat, co znamenají v analýze a co popisují. Obvykle jedna nebo dvě, maximálně však tři, nové souřadnicové osy *PC*1*, PC*2 a *PC*3 vystihnou dohromady největší podíl proměnlivosti dat.

První hlavní komponenta proložením nejmenšími čtverci. Účelem je proložit přímku v prostoru x_1 x_2 a x_3 umístěnými objekty. Když z každého *i-tého* objektu spustíme k této přímce kolmici, obdržíme vzdálenost ***e***, zvanou *reziduum.* Nejtěsněji proloženou přímku pak dostaneme, když suma čtverců reziduí bude dosahovat své minimální velikosti $RSS = \sum_{n} e_i^2 = \min$. Existují tedy dvě kritéria ke konstrukci první hlavní komponenty: 1. Směr největšího rozptylu, 2. metoda nejmenších čtverců. Ukazuje se, že obě metody vedou ke stejnému výsledku. Může nastat případ, že první hlavní komponenta *PC*1 nestačí dostatečně popsat rozptýlení objektů. Druhá hlavní komponenta *PC*2 je ortogonální vůči první (je na ni kolmá) a je umístěna do směru druhého největšího rozptýlení objektů v rovině. Podobně, vytvoří-li objekty v prostoru přibližný elipsoid, budou do prostorových os elipsoidu umístěny tři osy vystihující největší rozptýlení objektů čili osy tří hlavních komponent *PC*1*, PC*2 a *PC*3. Nové tři osy nekorelují, jsou vzájemně ortogonální, tj. na sebe kolmé. Metoda hlavních komponent PCA umožňuj e tuto geometrickou projekci zobecnit na libovolný počet *m* znaků.

Zaměření metody PCA: Poprvé byla zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem analýzy hlavních komponent je zjednodušení popisu skupiny vzájemně lineárně závislých neboli korelovaných znaků čili rozklad zdrojové matice dat do *matice strukturní* a do *matice šumové.* V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné nazvané *hlavní komponenty.* Každá hlavní komponenta představuje lineární kombinací původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl.* Hlavní komponenty jsou seřazeny dle důležitosti, tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.
Standardním využitím PCA je snížení dimenze úlohy čili redukce počtu znaků bez velké ztráty informace, a to užitím pouze prvních několika hlavních komponent. Toto snížení dimenze úlohy se netýká počtu původních znaků. Je výhodné především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními

vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných hlavních komponent.
Podstata metody PCA: Zdrojová matice dat $X(n \times m)$ obsahuje n objektů a m znaků. *Objekty* jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny atd., zatímco *znaky* či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický atd.), barva apod. Důležitá je zde skutečnost, že každý znak je znám pro všech n objektů. Správná skladba zdrojové matice $\boldsymbol{X}$ čili volba které znaky použít a které objekty zařadit je delikátní úkol silně závislý na charakteru každé úlohy. Cílem každé vícerozměrné analýzy je zpracovat data tak, aby se zřetelně indikoval model a tak odkryl skrytý jev.

Analýza shluků (CLU): Analýza shluků (Cluster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Navzdory starému přísloví, že opaky se přitahují, v přírodě se ukazuje, že platí spíše pravidlo, že podobné věci se sjednocují. Tato klasifikace se nazývá numerická taxonomie. V analýze shlukuje neznámá příslušnost do tříd všech objektů. Dokonce i počet tříd či shlukuje neznámý. 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.

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

Míry podobnosti: Myšlenka podobnosti objektů je v analýze shluků základní. 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. Meziobjektová podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin, a to míry korelace, míry vzdálenosti a míry asociace. Každá z nich představuje zvláštní pohled na podobnost, která je závislá na objektech a na typu dat. Korelační a

vzdálenostní míry jsou míry metrických dat, zatímco asociační míry jsou určeny spíše pro nemetrická data.

Korelační míry. Základní mírou podobnosti dvou objektů či znaků 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 jedné. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Obyčejně se vychází z transponované matice dat $\boldsymbol{X}^{\mathrm{T}}$, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice $\boldsymbol{X}^{\mathrm{T}}$ představují korelaci mezi dvojicí objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou „podobnost" a nízká korelace pak „nepodobnost" profilů. Korelační míry představují podobnost odpovídajících si „vzorů" posuzovanou přes všechny znaky. Korelační míry se však užívají zřídka, protože v praktické analýze je kladen důraz více na velikosti objektů než na tvar jejich profilových křivek.

Míry vzdálenosti. Představují nejčastěji užívané míry založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Pokud tyto míry splňují požadavky symetrie $d(x, y) = d(y, x)$ a trojúhelníkovou nerovnost $d(x, y) < d\{x, z) + d(y, z)$, jde o tzv. *metriky.* Při porovnání na profilovém diagramu je zřejmé, že vzdálenosti se zaměřují na velikost hodnot a vyhledání křivek v profilovém diagramu, které jsou blízko sebe, i když u jednotlivých znaků velmi odlišného tvaru. Použití korelace vede na jiné shluky než použití vzdálenostních měr. 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

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

představuje standardní typ vzdálenosti. Vedle eukleidovské vzdálenosti se užívá také *čtverec eukleidovské vzdálenosti,* který tvoří základ Wardovy metody shlukování. Existuje však ještě několik dalších vzdálenostních měr.

Dendrogramy hierarchického shlukování: Analýzou shlukuje 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, když je k dispozici matice původních znaků.

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 schopnost.

Míra věrohodností: Při sestrojování dendrogramu, k čemuž máme celou řadu technik, posuzujeme *míru věrohodnosti* nebo-li *těsnost proložení.* Prvním kritériem při volbě „nejlepšího dendrogramu", jenž 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. Čím vyšší hodnota *CC,* tím větší věrohodnost a lepší model shluků. Druhým kritériem těsnosti proložení je *kritérium delta* A, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti. Je žádoucí, aby hodnoty kritéria delta byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Klasifikace zdrojů pitné vody

Na 62 vzorcích pitné vody, u kterých bylo provedeno stanovení 16 znaků kvality, proveďte vícerozměrnou statistickou analýzu. Je třeba provést exploratorní (průzkumovou) analýzu dat pro (a) posouzení *podobnosti objektů* pomocí rozptylových a symbolových grafů, (b) nalezení *vybočujících objektů*, resp. jejich znaků. Dále třeba zjistit, který původní znak nejvíce přispívá do první hlavní komponenty y_1 a který do druhé y_2? Který původní znak je v grafu komponentních vah nejdůležitější a který je nejméně důležitý - redundandní? Ukazuje tento graf na nějaké korelující znaky? Indikuje rozptylový diagram komponentního skóre nějaké odlehlé objekty-vzorky pitné vody? Kolik shluků objektů odhaluje dendrogram podobnosti objektů za použití optimální shlukovací procedury? Lze nalézt i dominantní znaky, které je třeba měřit v analytické laboratoři a redundantní znaky, které je možné vypustit? Shlukovou analýzou proveďte klasifikaci vzorků pitné vody. Jsou závěry metod PCA a FA odlišné?

Data: Výběr se týká 62 vzorků v řádcích a 16 sledovaných znaků ve sloupcích a obsahuje *i* (ID) pořadové číslo vzorku a *x1* (NO3) značí obsah dusičnanů [mg/l], *x2* (NO2) značí obsah dusitanů [mg/l], *x3* (CL) značí obsah chloridů [mg/l], *x4* (CL2) značí obsah celkového chloru [mg/l], *x5* (SO4) značí obsah síranů [mg/l], *x6* (PO4) značí obsah fosforečnanů [mg/l], *x7* (NH4) značí obsah amonných solí [mg/l], *x8* (CA) značí obsah vápníku [mg/l], *x9* (MG) značí obsah hořčíku [mg/l], *x10* (FE) značí obsah železa (celkového) [mg/l], *x11* (MN) značí obsah manganu [mg/l], *x12* (PH) značí hodnotu pH, *x13* (KNK) značí hodnotu KNK, *x14* (ZNK) značí hodnotu ZNK, *x15* (VODIV) značí vodivost, *x16* (NEROZP) značí nerozpuštěné látky [mg/l].

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*	*x15*	*x16*
1	2.2	0.00	6	6.	103.5	0.032	0.02	181	17	0.016	0.05	7.08	8.1	3.40	855	0.09
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
62	32.8	0.01	25	25.	115.5	0.050	0.02	102	12	0.016	0.05	7.69	2.6	0.65	436	0.05

Řešení: přináší obrázky ve slidech.

Poděkování: Práce vznikla za podpory grantu Ministerstva zdravotnictví NS9831-4/2008 a vědeckých záměrů Ministerstva školství, kultury a mládeže MSMT0021627502.

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.
[3] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.
[4] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[5] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.
[6] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[7] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis,* Prentice Hall, 1982
[9] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.
[10] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[11] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[12] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004..
[13] Meloun M. , Militký J., Hill M., *Počítačová analýza vícerozměrných dat v příkladech,* Academia Praha 2005.
[15] Meloun M. , Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002, 2006.

Odhalení skryté struktury a vnitřních vazeb dat vícerozměrnou statistickou analýzou pitné vody

Prof. RNDr. Milan Meloun, DrSc,
Katedra analytické chemie,
Univerzita Pardubice, 532 10 Pardubice,

milan. meloun@upce. cz

Klasifikace zdrojů pitné vody

Zadání úlohy: Na 62 vzorcích pitné vody bylo provedeno stanovení 16 znaků kvality, proveďte vícerozměrnou statistickou analýzu.

Data: Výběr má 62 vzorků v řádcích a 16 sledovaných znaků ve sloupcích:
NO3 značí obsah dusičnanů [mg/l],
NO2 značí obsah dusitanů [mg/l],
CL značí obsah chloridů [mg/l],
CL2 značí obsah celkového chloru [mg/l],
SO4 značí obsah síranů [mg/l],
PO4 značí obsah fosforečnanů [mg/l],
NH4 značí obsah amonných solí [mg/l],
CA značí obsah vápníku [mg/l],
MG značí obsah hořčíku [mg/l],
FE značí obsah železa (celkového) [mg/l],
MN značí obsah manganu [mg/l],
PH značí hodnotu pH,
KNK značí hodnotu KNK,
ZNK značí hodnotu ZNK,
VODIV značí vodivost,
NEROZP značí nerozpuštěné látky [mg/l].

Postup analýzy vícerozměrných dat

1. Standardizace: analýze vždy předchází standardizace čili škálování proměnných.

2. Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:

Vyčíslení výběrové střední hodnoty každé proměnné.
Odhad kovarianční matice S a její normované podoby - korelační matice R.
Odhadu vícerozměrné šikmosti a vícerozměrné špičatosti.
Matice R obsahuje Pearsonovy párové korelační koeficienty, které se diskutují.

3. Exploratorní analýza dat EDA:

(a) Hledání podobnosti objektů vizuálními rozptylovými diagramy typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy),

(b) Nalezení vybočujících objektů nebo vybočujících proměnných, mnohdy nevhodných k analýze,

(c) Testy předpokladů lineárních vazeb,

(d) Testy předpokladů o datech (normalitu, nekorelovanost, homogenitu).
Ověřování normality založené na vícerozměrné šikmosti a vícerozměrné špičatosti.

4. Určení vhodného počtu latentních proměnných:

a) Matice S nebo R se rozloží na vlastní čísla a vlastní vektory.

b) Indexový graf úpatí vlastních čísel (Scree plot): určí vhodný počet latentních proměnných, které ještě dostatečně popisují proměnlivost v datech.

c) Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. Jinak jde o hlavní komponenty.

5. Určení struktury v proměnných (PCA a FA):

Graf komponentních vah (Plot of components weights, loadings): hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu

a) Rozptylový diagram komponentního skóre (Scatterplot): hledání struktury v objektech a třídění objektů do shluků.

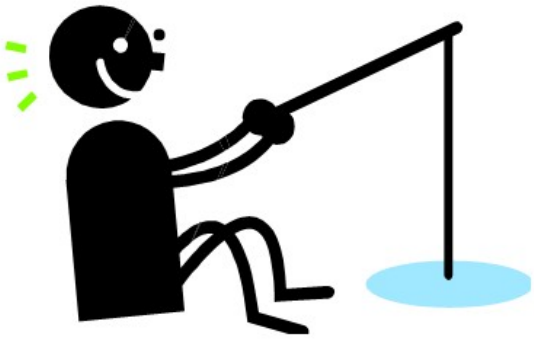
b) Dvojný graf (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

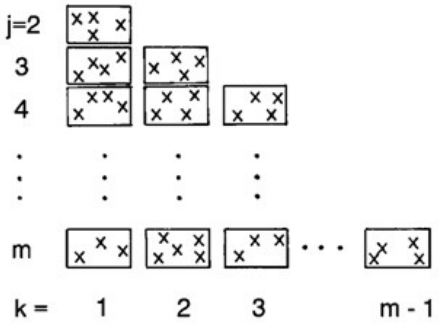
Zobecněné rozptylové diagramy a symbolové grafy

exploratorní analýzy vícerozměrných dat

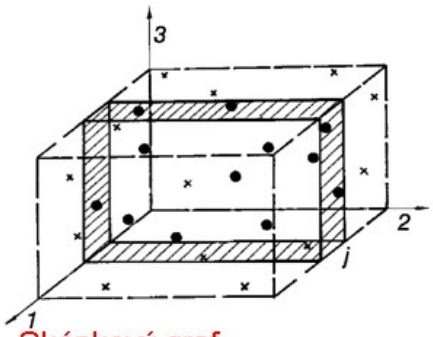
(a) Hledání podobnosti objektů vizuálními rozptylovými diagramy typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy),

(b) Nalezení vybočujících objektů nebo vybočujících proměnných, mnohdy nevhodných k analýze,

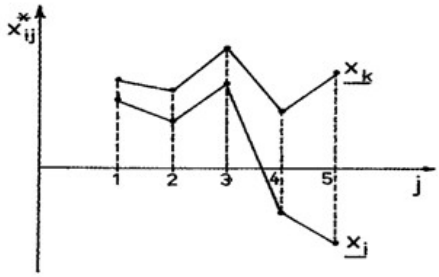




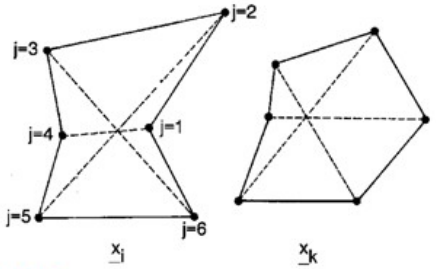
Párové rozptylové diagramy



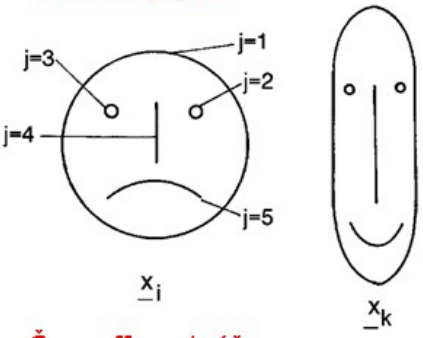
Okénkový graf



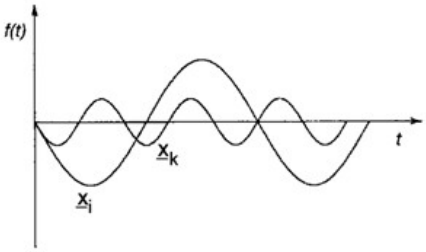
Profily



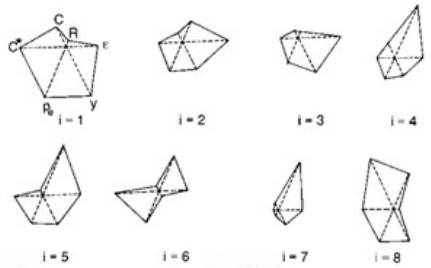
Polygony



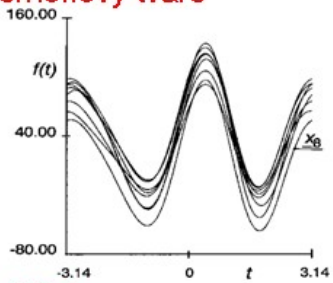
Černoffovy tváře



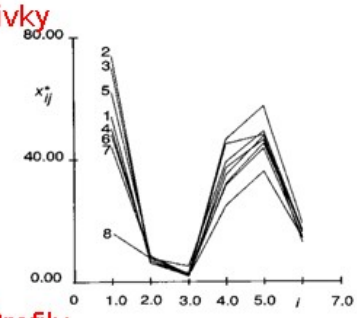
Křivky



Polygony – hvězdičky



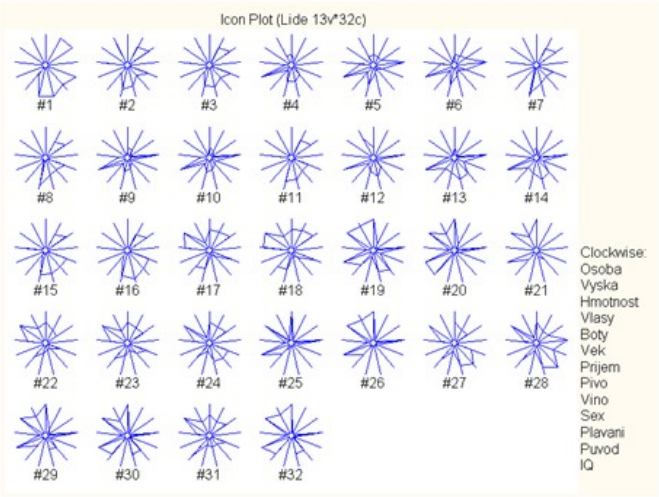
Křivky



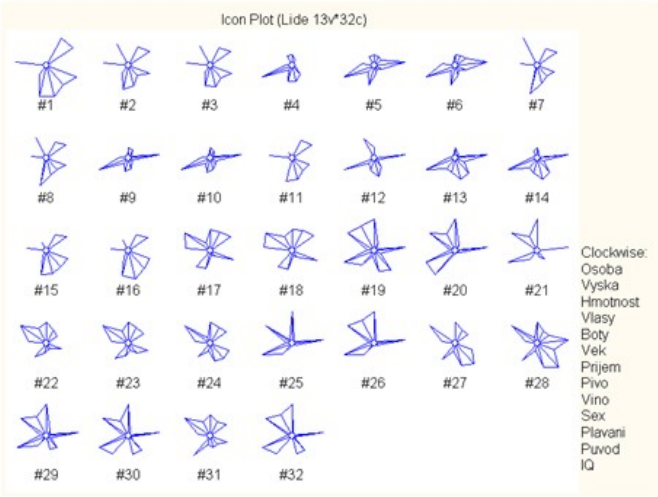
Profily

Zobecněné rozptylové grafy a symbolové grafy

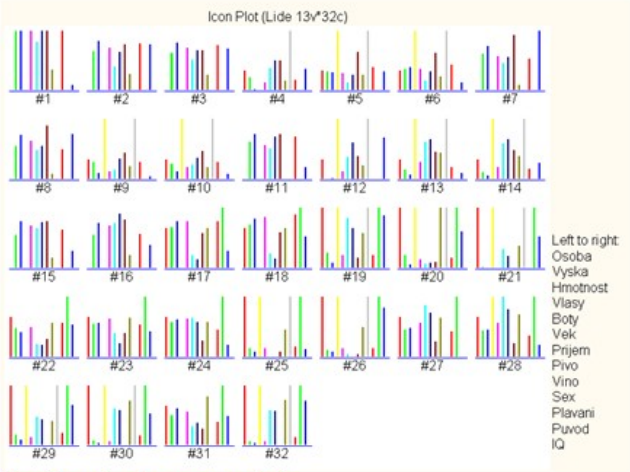
exploratorní analýzy vícerozměrných dat datové matice



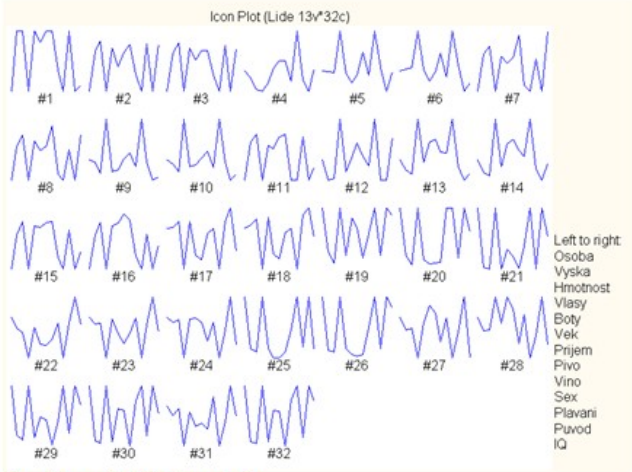
Polygony – sluníčka dat



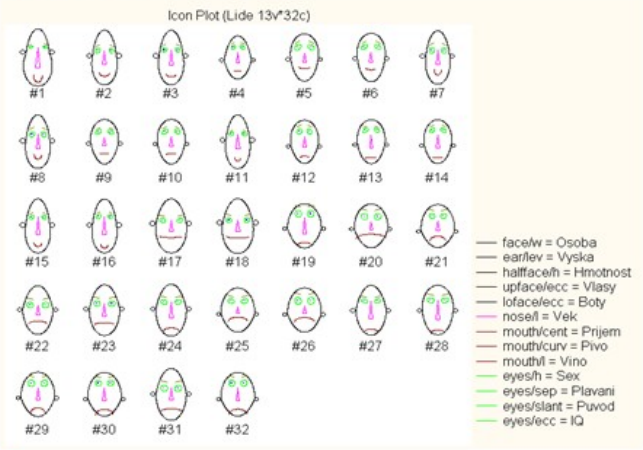
Polygony – hvězdičky dat



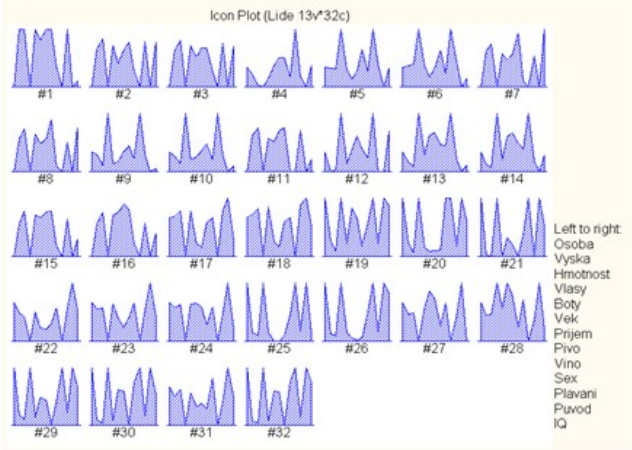
Profilové sloupce dat



Profilové křivky dat

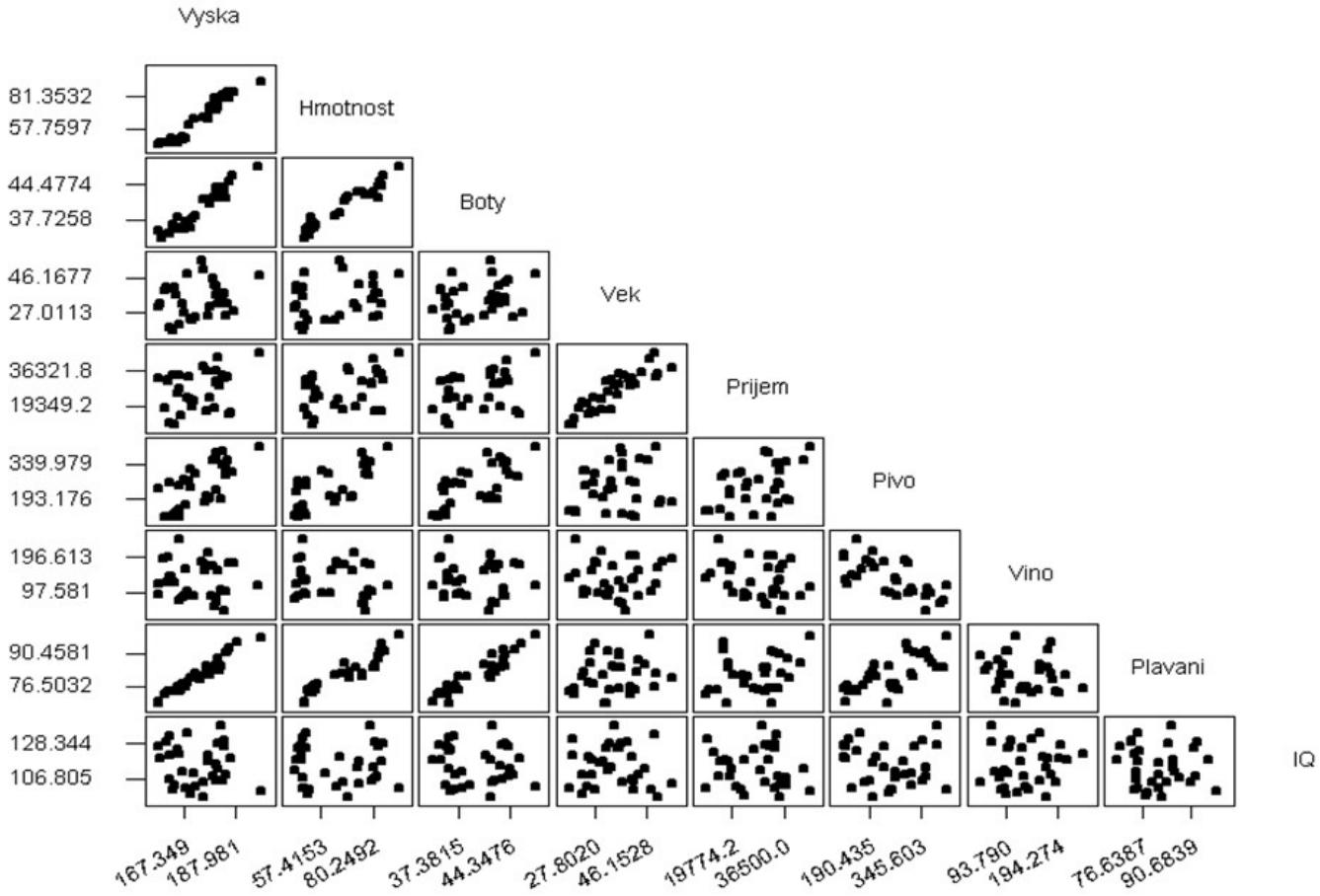


Černoffovy tváře dat



Profily znaků dat

Ukázka grafického zobrazení korelační matice dat

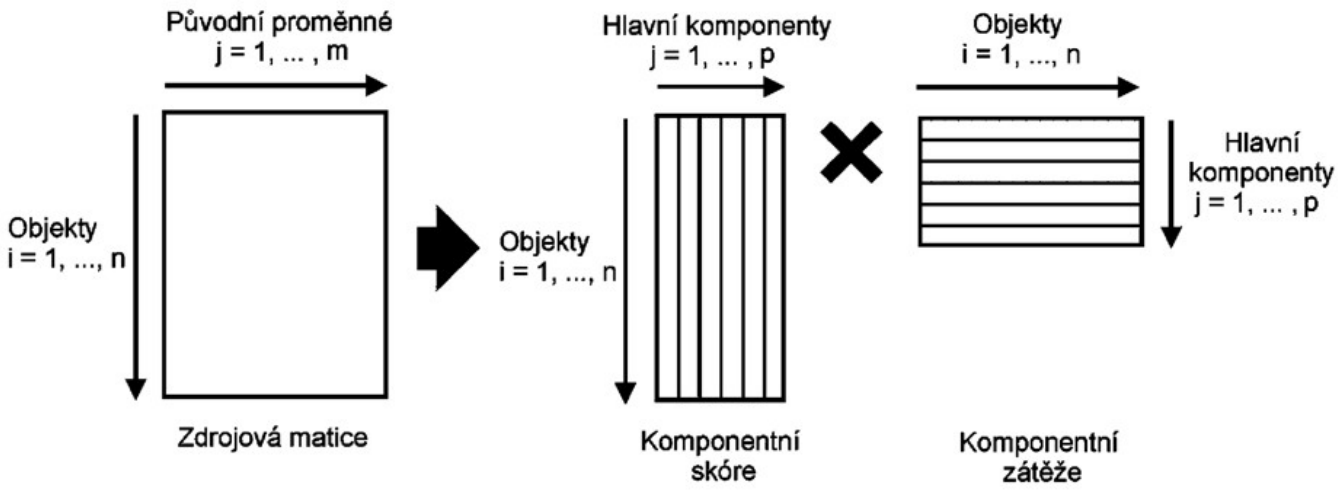


Ukázka korelační matice dat u úlohy pitné vody

Červeně jsou značeny statisticky významné hodnoty korelačního koeficientu.

	NO3	NO2	Cl	Cl2	SO4	PO4	NH4	Ca	Mg	Fe	Mn	pH	KNK	ZNK	Vodiv	Neroz
NO3	1.000000															
NO2	0.047190	1.000000														
Cl	0.075647	0.040778	1.000000													
Cl2	0.081276	0.042804	**0.998609**	1.000000												
SO4	-0.053931	-0.126076	**0.456812**	**0.466434**	1.000000											
PO4	-0.020737	**0.465180**	0.032834	0.033436	-0.197464	1.000000										
NH4	-0.109845	-0.091387	-0.13722	-0.139016	0.050398	-0.055856	1.000000									
Ca	-0.017772	-0.175622	**0.539559**	**0.550111**	**0.755038**	-0.075953	0.060288	1.000000								
Mg	-0.123426	-0.186858	**0.476512**	**0.473684**	**0.381889**	-0.138298	-0.082867	0.155819	1.000000							
Fe	-0.085330	0.049748	0.012382	0.013248	**0.316056**	0.028380	-0.075350	0.082685	**0.369682**	1.000000						
Mn	-0.215441	-0.042447	-0.04827	-0.047532	0.042870	-0.013695	**0.399794**	0.028817	-0.157013	-0.048317	1.000000					
pH	0.157820	**-0.358610**	-0.06209	-0.051200	**-0.283529**	-0.056366	-0.215715	**-0.285059**	-0.019601	-0.106596	-0.249614	1.000000				
KNK	-0.232680	-0.157834	**0.406614**	**0.389809**	**0.265703**	0.028287	0.094545	**0.515586**	**0.495420**	0.173567	-0.126195	-0.297318	1.000000			
ZNK	-0.226557	-0.067831	**0.344606**	**0.331544**	**0.441636**	0.205221	0.091070	**0.654338**	**0.310340**	0.143989	0.004198	-0.411754	0.759789	1.000000		
Vodiv	-0.050967	-0.106941	**0.648153**	**0.621040**	**0.515015**	-0.046916	0.027953	**0.609324**	**0.449503**	0.185218	-0.063259	-0.342510	0.733499	0.693455	1.000000	
Nerozp	-0.047249	-0.055412	0.072002	0.067766	-0.054792	-0.126639	0.005071	0.072844	0.021429	0.137923	-0.050188	-0.090398	0.219339	-0.103607	0.253084	1.000000

Metoda hlavních komponent (PCA)



Původní zdrojová matice dat se rozloží na matici komponentního skóre a matici komponentních zátěží.

Diagnostika metody PCA

Maticový graf rozptylových diagramů znaků slouží k získání počáteční informace o datech, zda data potřebují škálování. V PCA postupně provádíme:

1. ***Vyšetření indexového grafu úpatí vlastních čísel*** – z hrany úpatí v tomto diagramu se určí vhodný počet hlavních komponent.

2. ***Výpočet vlastních vektorů*** – vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních vektorů, který přehledně informuje o relativním zastoupení původních znaků x_j , $j = 1, ..., m$, v hlavních komponentách.

3. ***Výpočet komponentních vah*** – matice párových korelačních koeficientů obsahující korelace původních znaků s hlavními komponentami. Uživatel nyní vybere pouze prvních k hlavních komponent a vytvoří tak model PCA.

4. ***Vyšetření grafu komponentních vah.***

5. ***Vyšetření rozptylového diagramu komponentního skóre.***

6. ***Vyšetření dvojného grafu.***

7. ***Vyšetření reziduí*** – rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení.

8. ***Určení významných původních znaků*** – je výhodné vyhledávat významné znaky, protože klasická metoda PCA umožňuje sice redukci počtu hlavních komponent, ale každá komponenta zůstává stále kombinací všech původních znaků.

Výklad:

Indexový graf úpatí vlastních čísel

(Scree Plot)

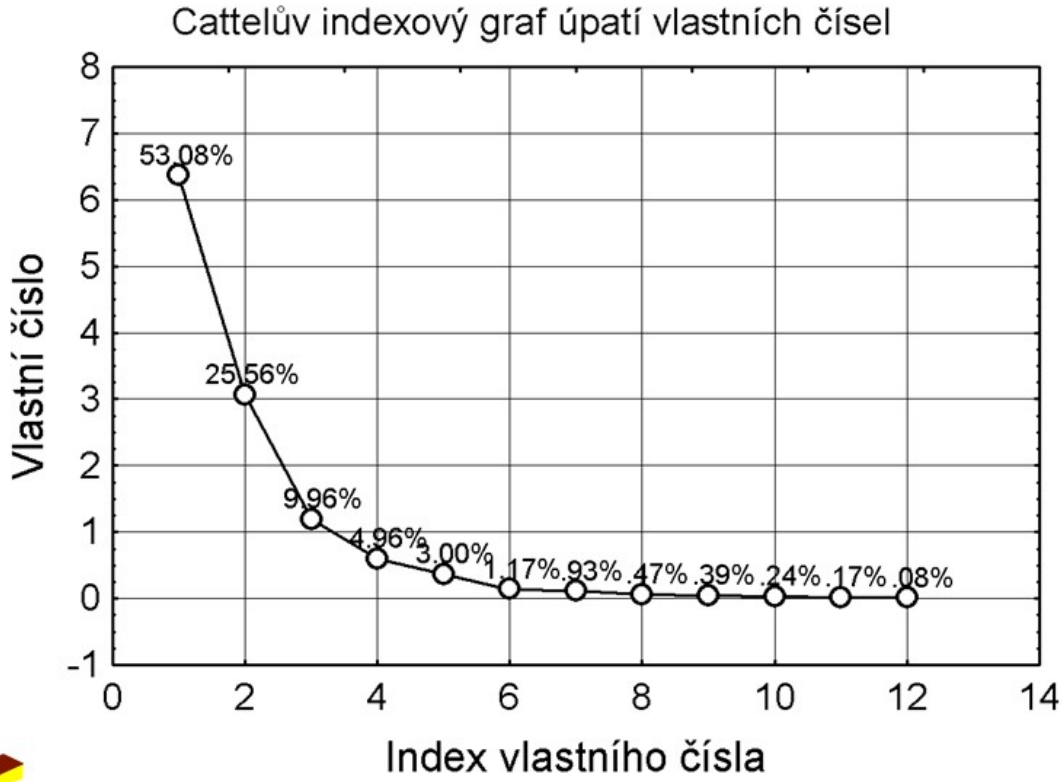
Je to sloupcový diagram vlastních čísel proti indexu *A*.

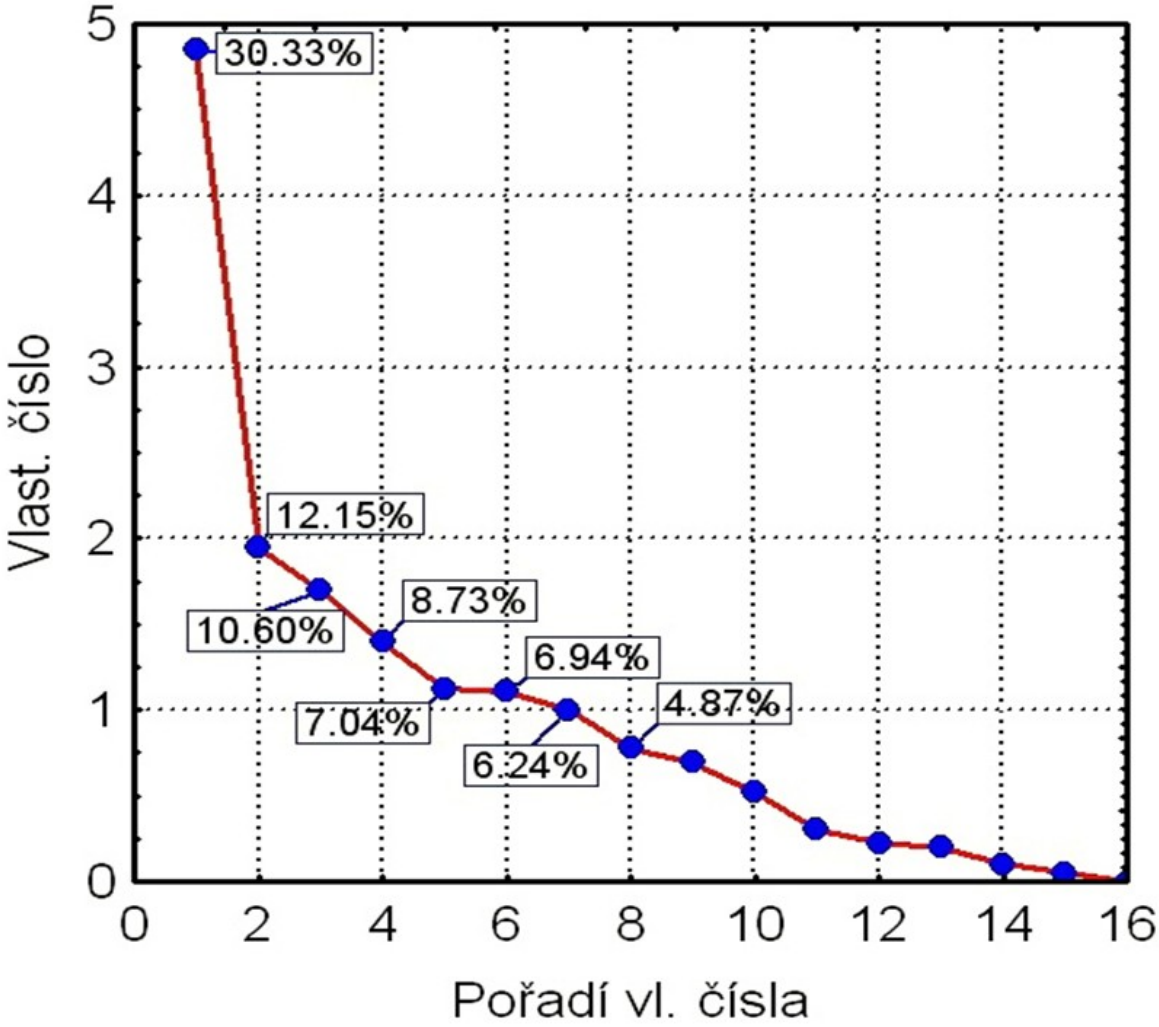
Zobrazuje: relativní velikost jednotlivých vlastních čísel.

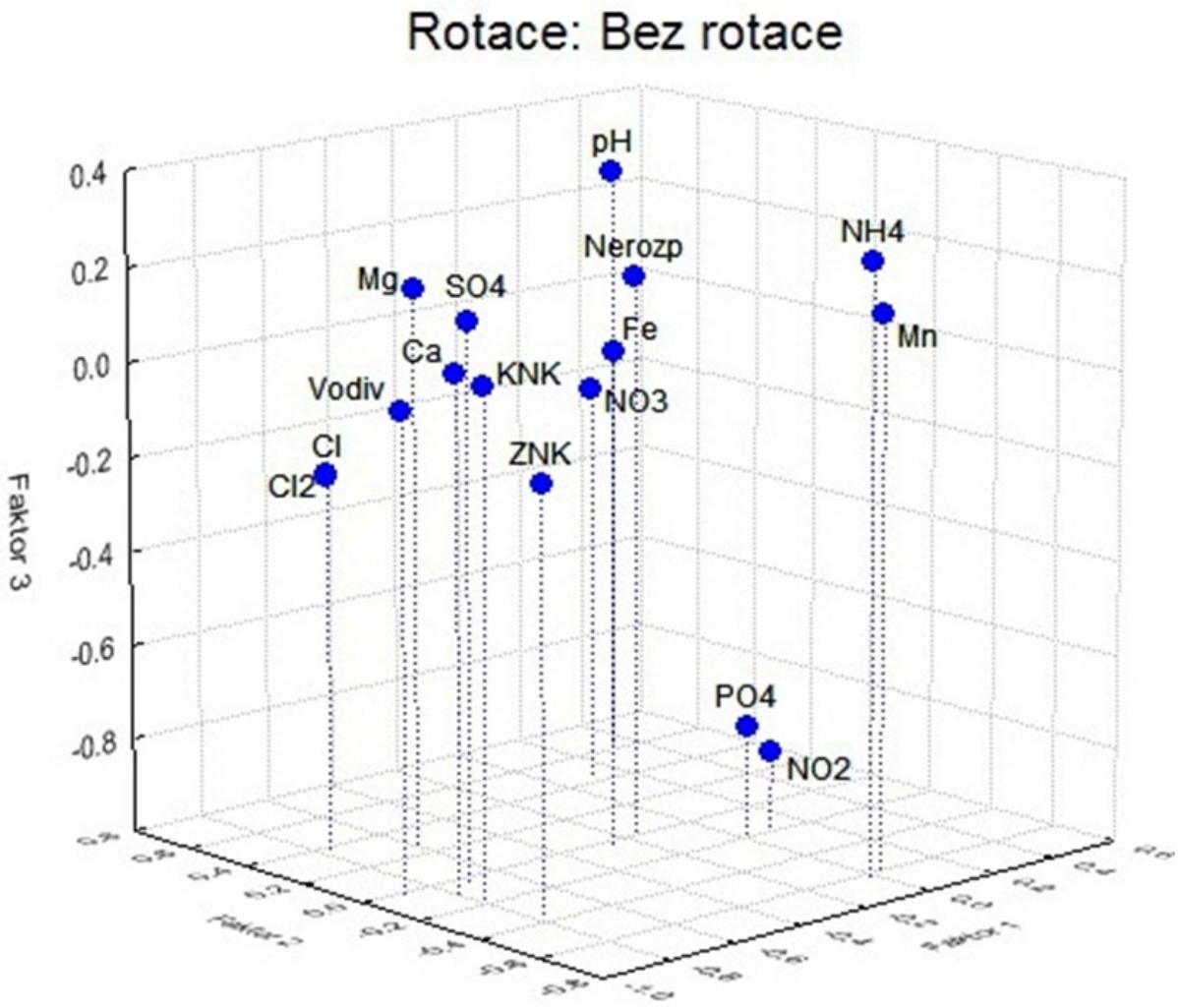
Využití: k určení počtu *A* "užitečných" hlavních komponent. Graf úpatí se jeví nejobjektivnějším kritériem.

Kritérium "1": hrubším kritériem PC, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.

Cattelův indexový graf úpatí vlastních čísel k určení počtu hlavních komponent







Výklad:

Graf komponentních vah, zátěží

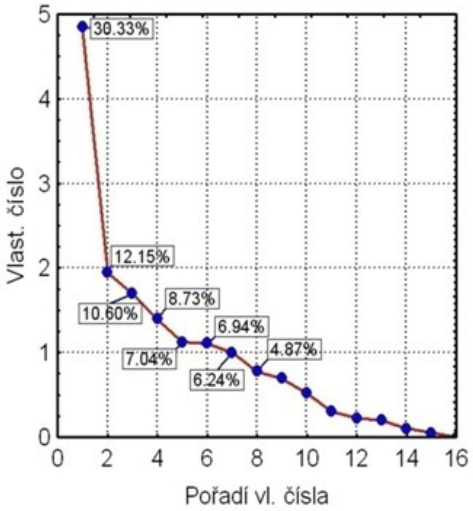
(Plot Components Weights)

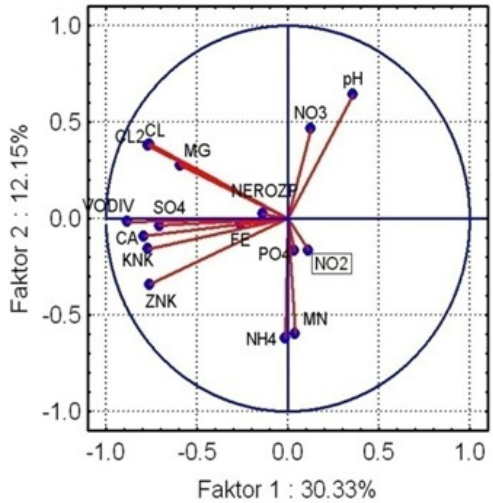
Zobrazí: komponentní váhy

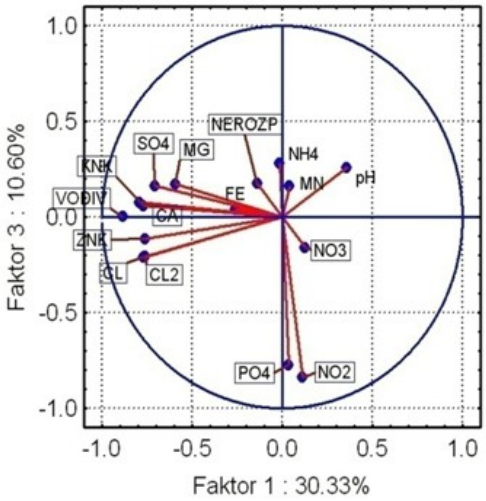
Porovnávají se: vzdálenosti mezi proměnnými.
Krátká znamená silnou korelaci.

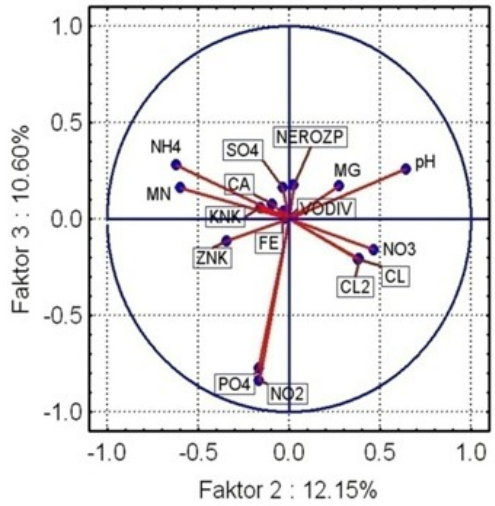
Nalezneme: shluk podobných proměnných, jež spolu korelují.

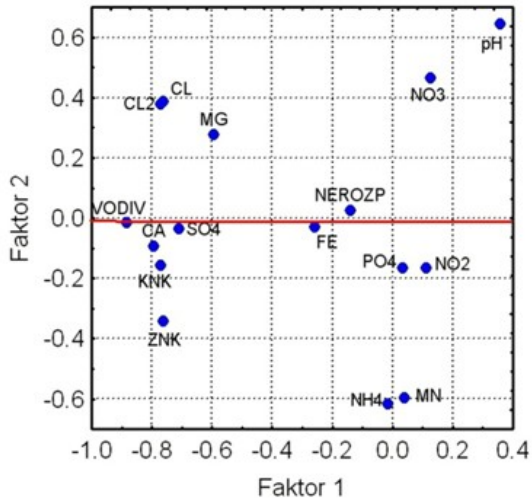
Představuje: most mezi původními proměnnými a hlavními komponentami.



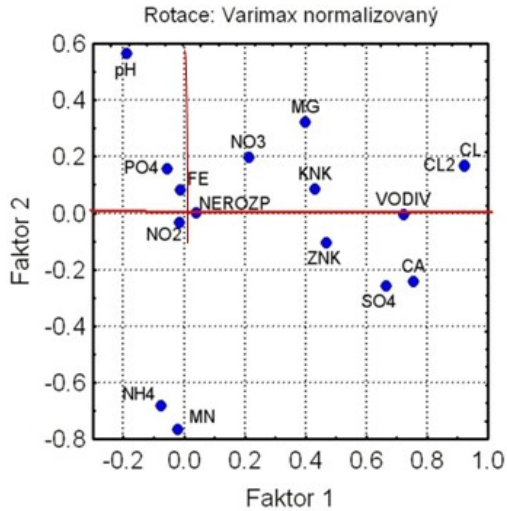


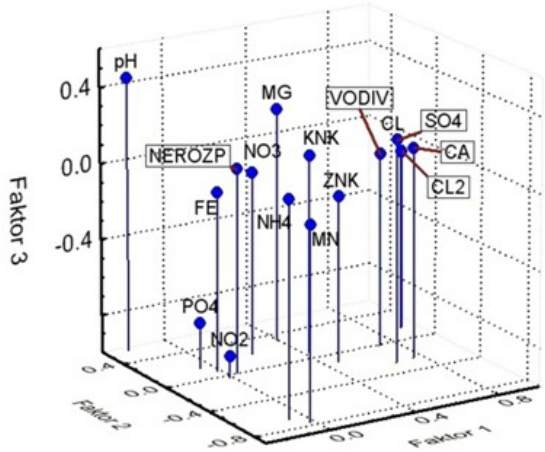




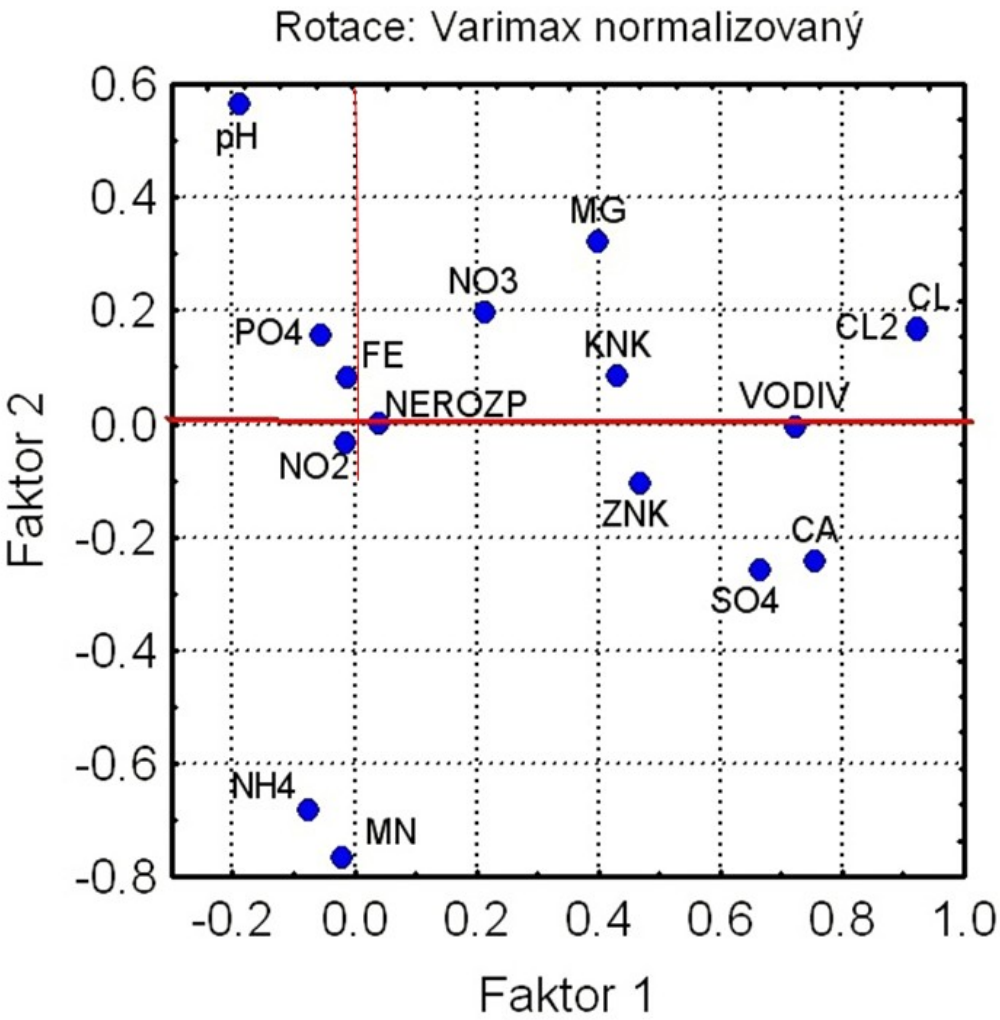


Rotace: Varimax normalizovaný





Cílem je vytvoření shluků podobných či vzájemně korelujících znaků ve 3D pros tří faktorů po rotaci, kde budou také zna faktorově čisté.



Rozptylový diagram komponentního skóre

(Scatterplot)

1. Umístění objektů: daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.

2. Podobnost objektů: objekty blízko sebe si jsou podobné, daleko od sebe jsou si nepodobné.

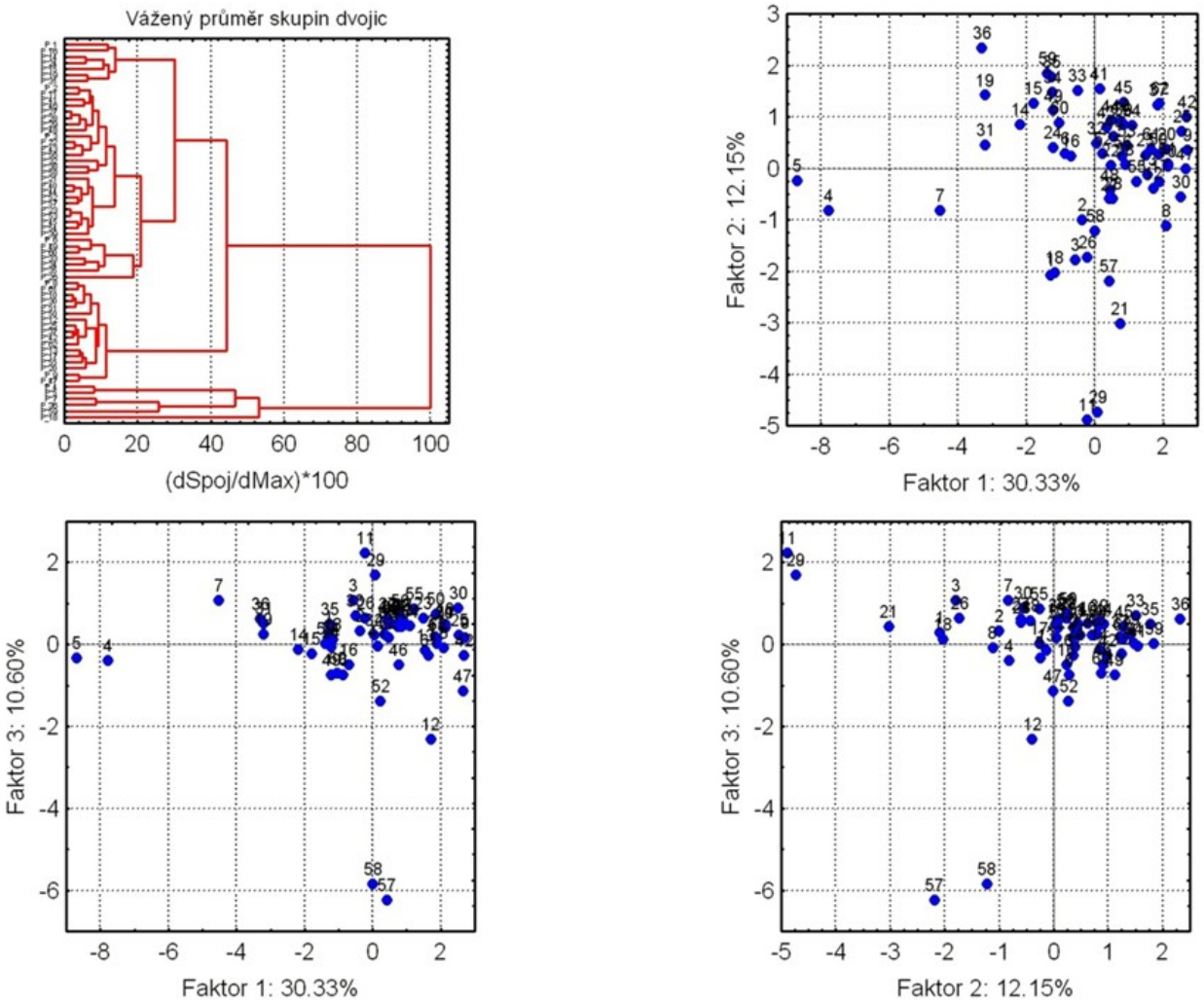
3. Objekty v shluku: umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.

4. Osamělé objekty: izolované objekty mohou být odlehlé.

5. Odlehlé objekty: ideálně bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je špatný model.

6. Pojmenování objektů: výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a vystihneme tak jejich fyzikální či biologický vztah.

7. Vysvětlení místa objektu: umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu.



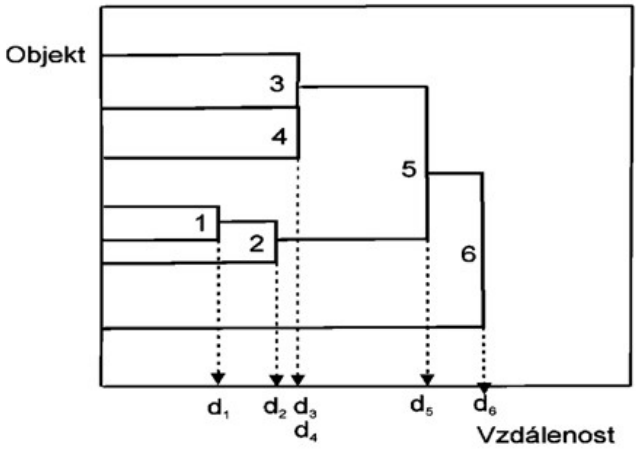
Hierarchické shlukování

konstrukce stromovité struktury, dendrogramu

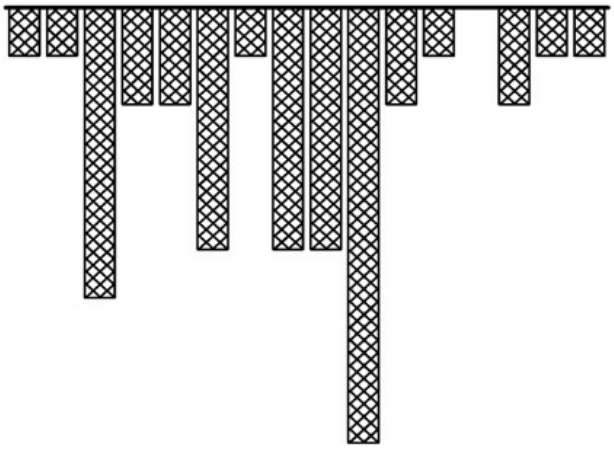
Způsoby hierarchického shlukování: aglomerační a divizní,

Aglomerační způsob: nejprve se spojí dva nejbližší objekty v jediný shluk, pak se připojí třetí objekt k prvním dvěma objektům a vznikne společný shluk. Tak se seskupí všechny objekty do jednoho velikého shluku.

1) růstový strom (dendrogram),



2) vertikální krápníkovitý diagram,



Aglomerační způsoby (algoritmy)

výstavby dendrogramu shluků:

Metoda nejbližšího souseda: je postavena na minimální vzdálenosti objektů.

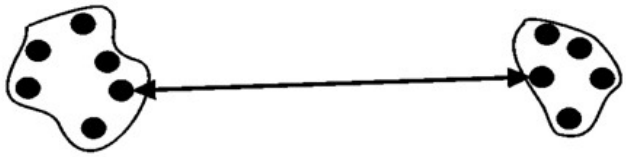
Metoda nejvzdálenějšího souseda: je postavena nikoliv na minimální ale na maximální vzdálenosti.

Metoda průměrového linkování: kritériem je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

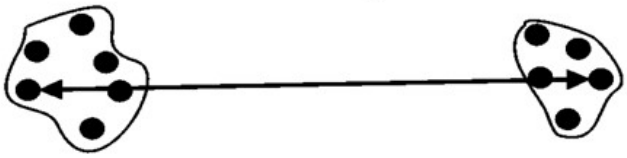
Wardova metoda: vzdálenost mezi dvěma shluky je tvořena na základě sumy čtverců přes všechny proměnné mezi dvěma shluky.

Metoda těžiště: vzdálenost těžišť shluků spojených Euklidovskou vzdáleností nebo čtvercem Euklidovské vzdálenosti. Těžiště shluku je průměrná hodnota objektů v proměnných, vyjádřená ve shlukových proměnných.

Metoda nejblizšího souseda

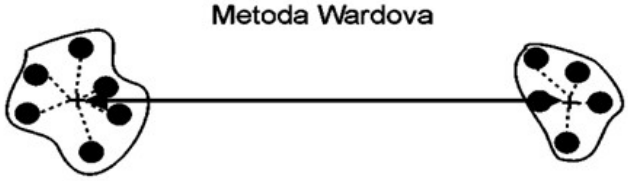


Metoda nevzdálenějšího souseda



Metoda průměrné vzdálenosti

Metoda Wardova



Míra věrohodnosti "nejlepšího dendrogramu"

1. kritérium těsnost proložení:

kofenetický korelační koeficient CC

- nejlépe odpovídá struktuře objektů a znaků mezi objekty,
- je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

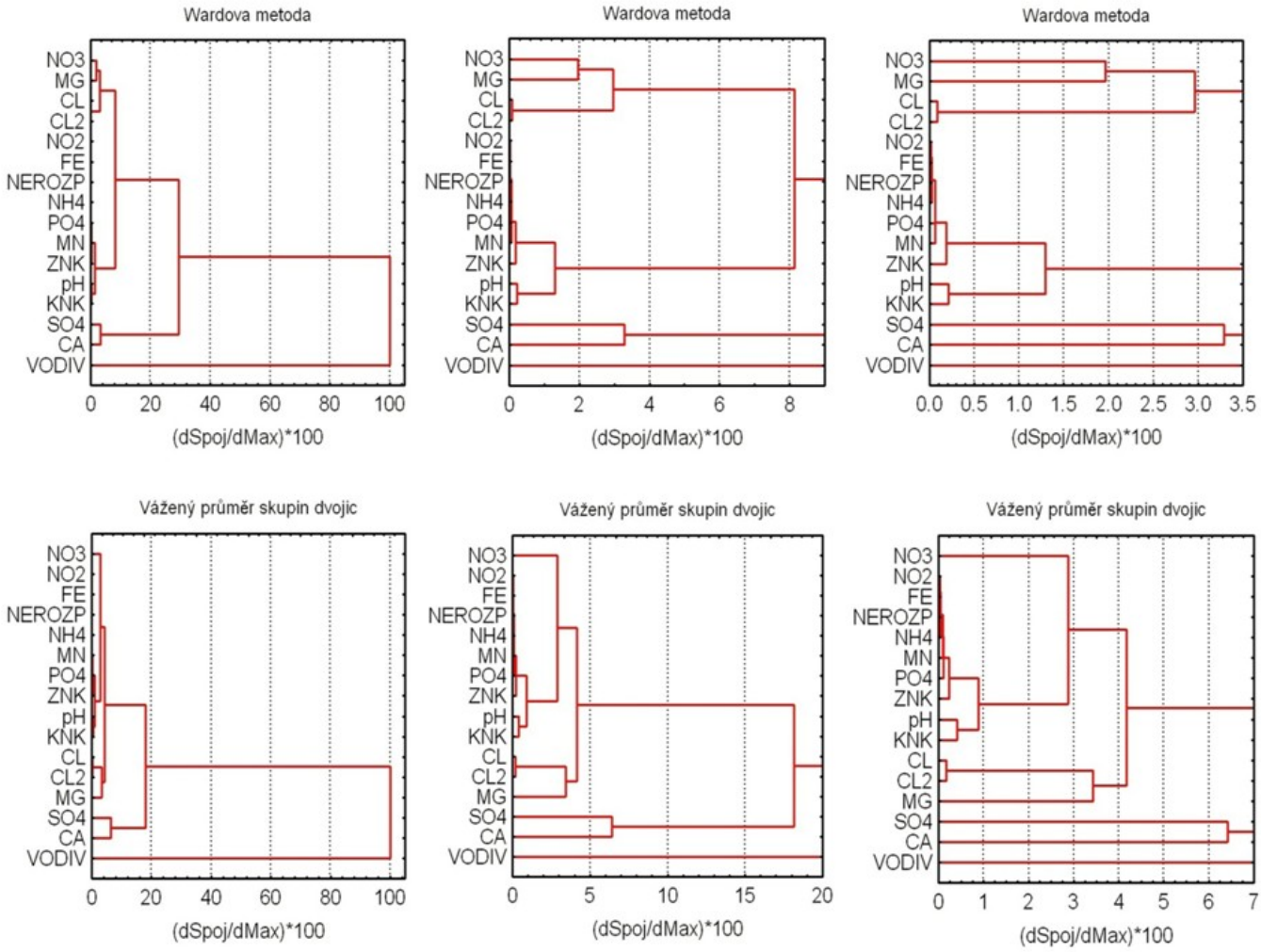
2. kritérium těsnosti proložení:

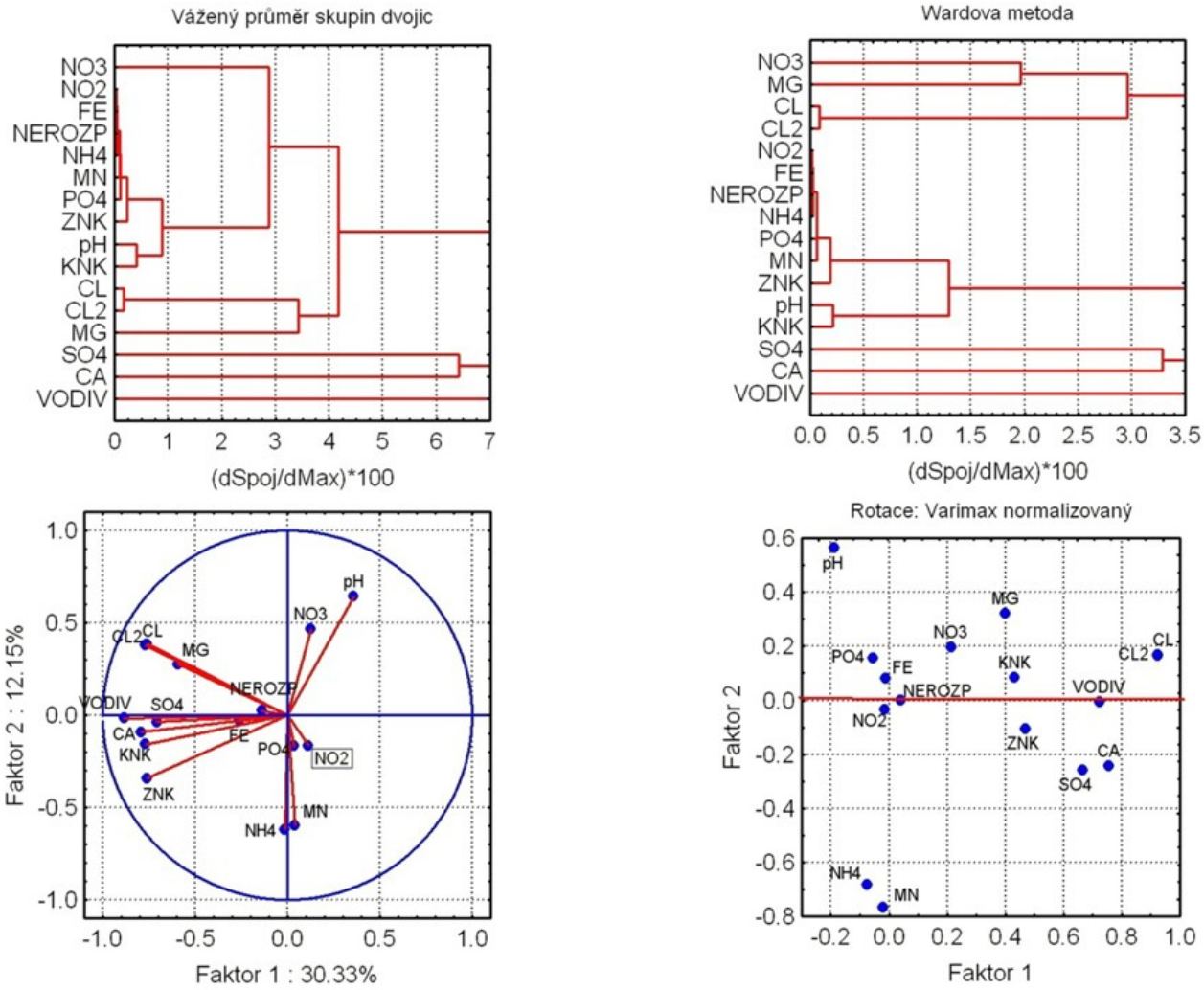
kritérium delta Δ

- měří stupeň přetvoření struktury dat,
- je žádoucí, aby hodnoty *delta* byly blízké nule,
- je definováno

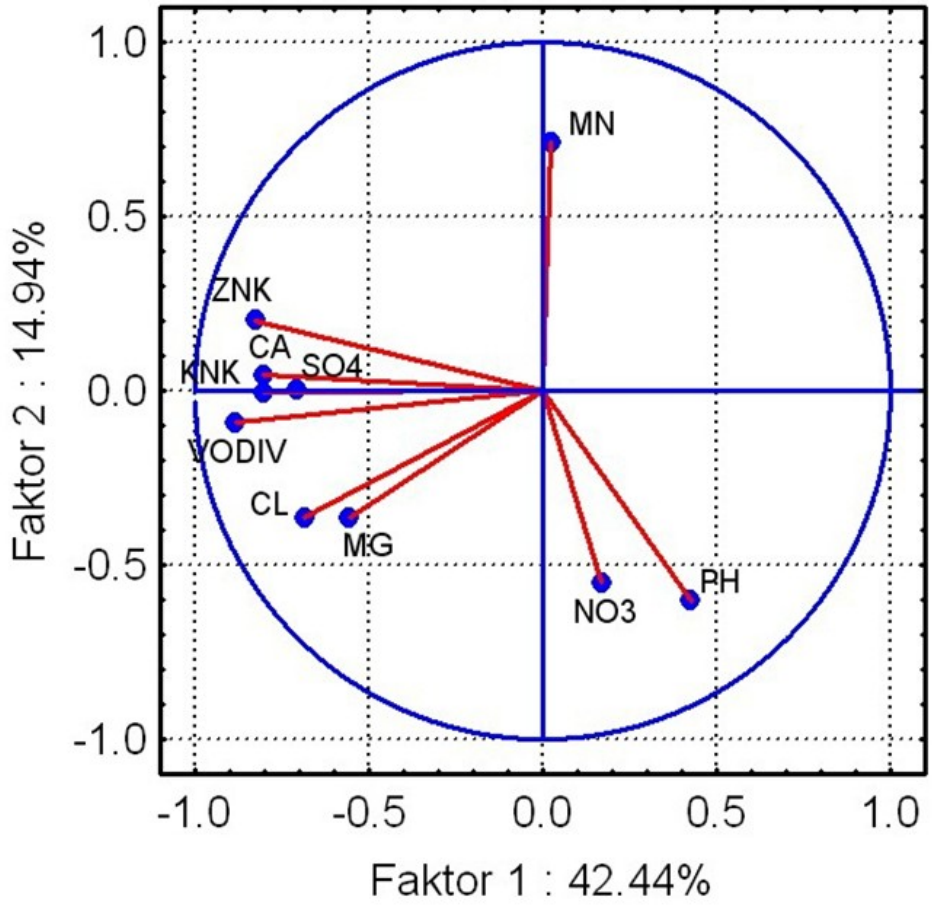
$$\Delta_A = \left[\frac{\sum_{j<k}^{N} \left| d_{jk} - d^*_{jk} \right|^{1/A}}{\sum_{j<k}^{N} \left(d^*_{jk} \right)^{1/A}} \right]^A$$

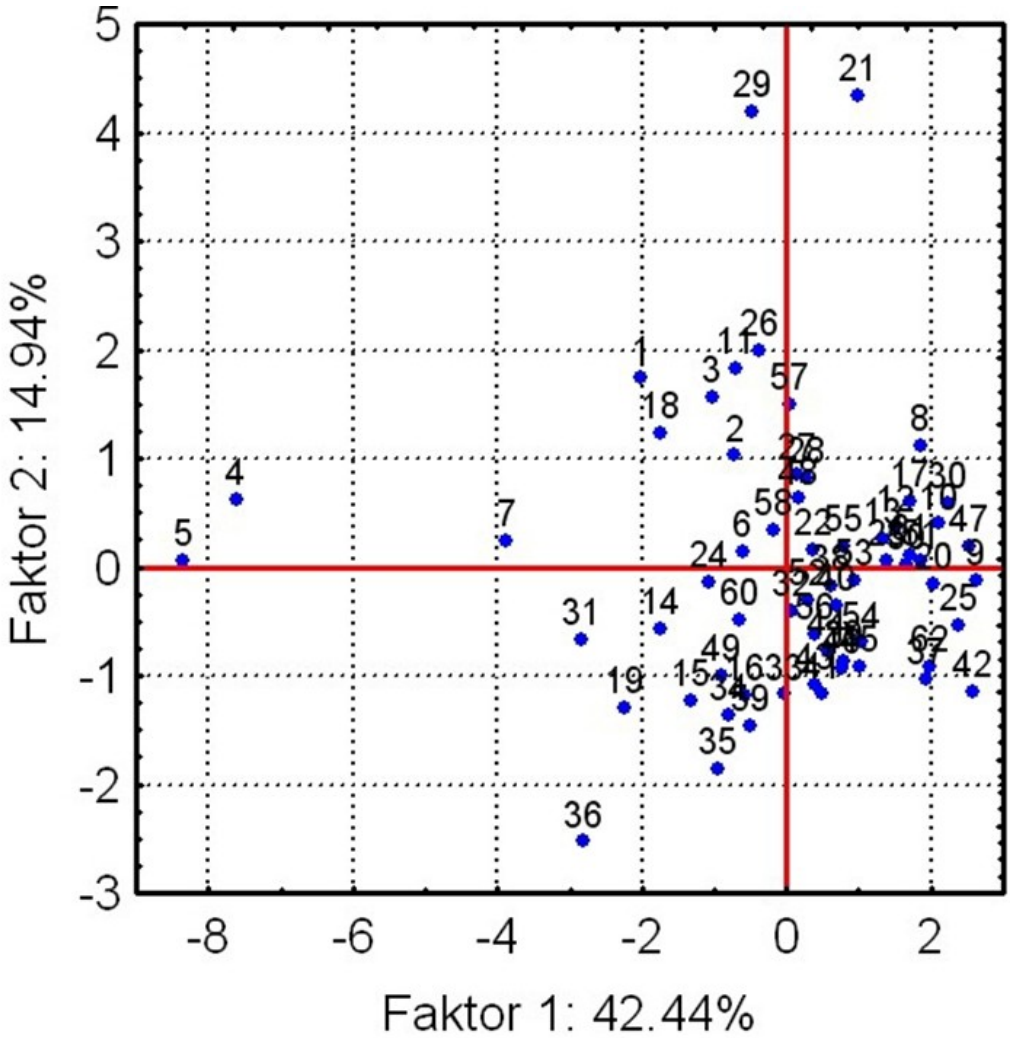
kde A = 0.5 nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu.

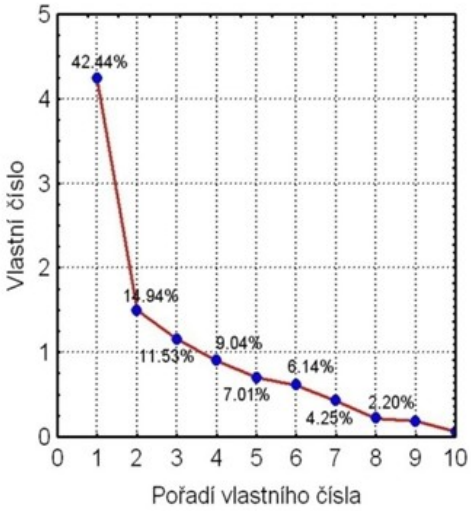


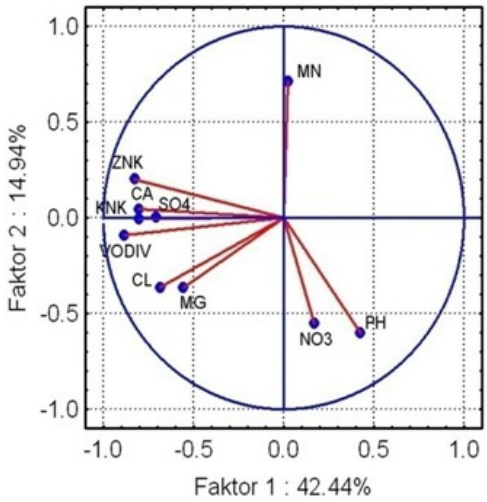


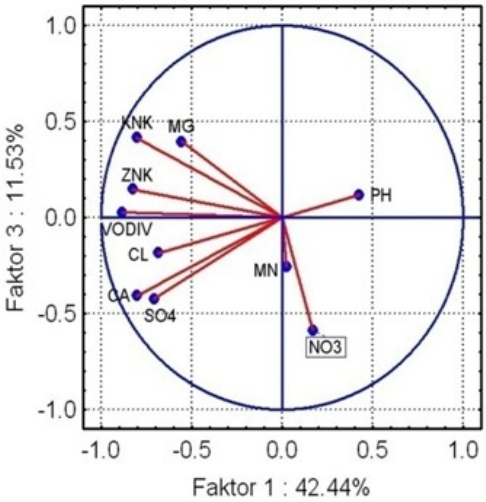
Po vypuštění redundantních znaků FE, NO2, PO4, CL2, NH4, Nerozp
které minimálně ovlivňují výběr:

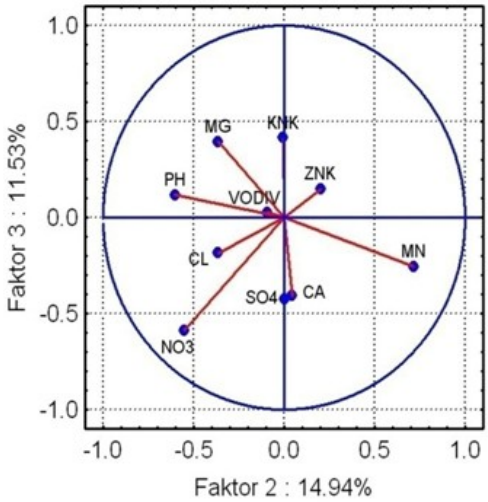


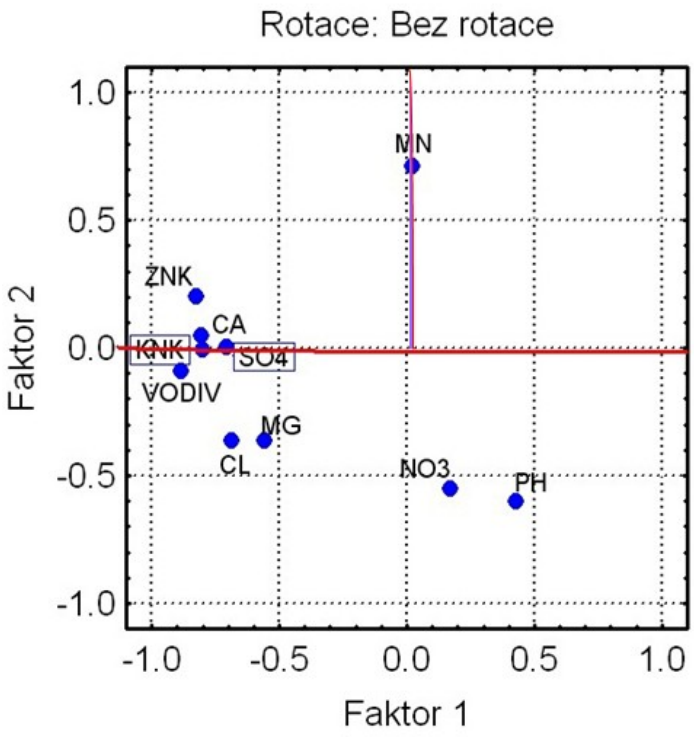

Faktor 2: 14.94%
Faktor 1: 42.44%

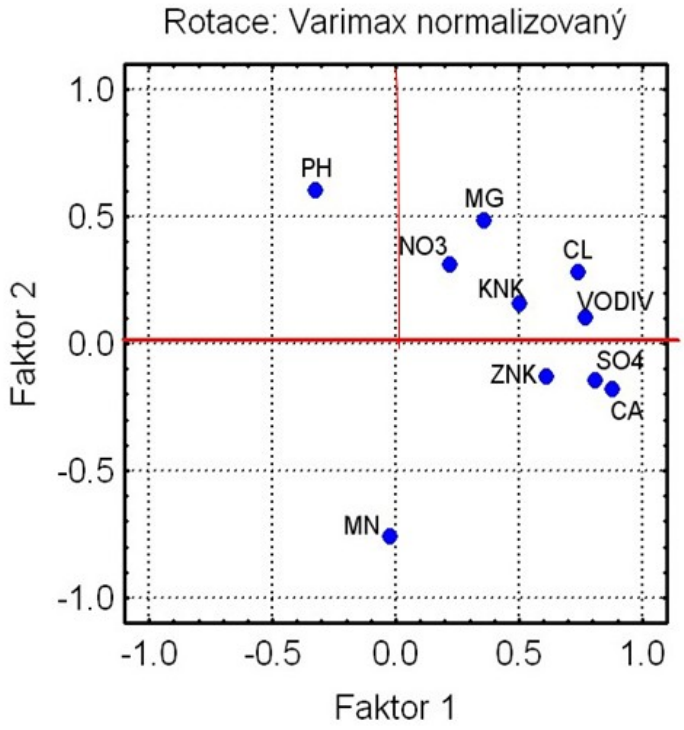

Vlastní číslo
Pořadí vlastního čísla
42.44%
14.94%
11.53%
9.04%
7.01%
6.14%
4.25%
2.20%

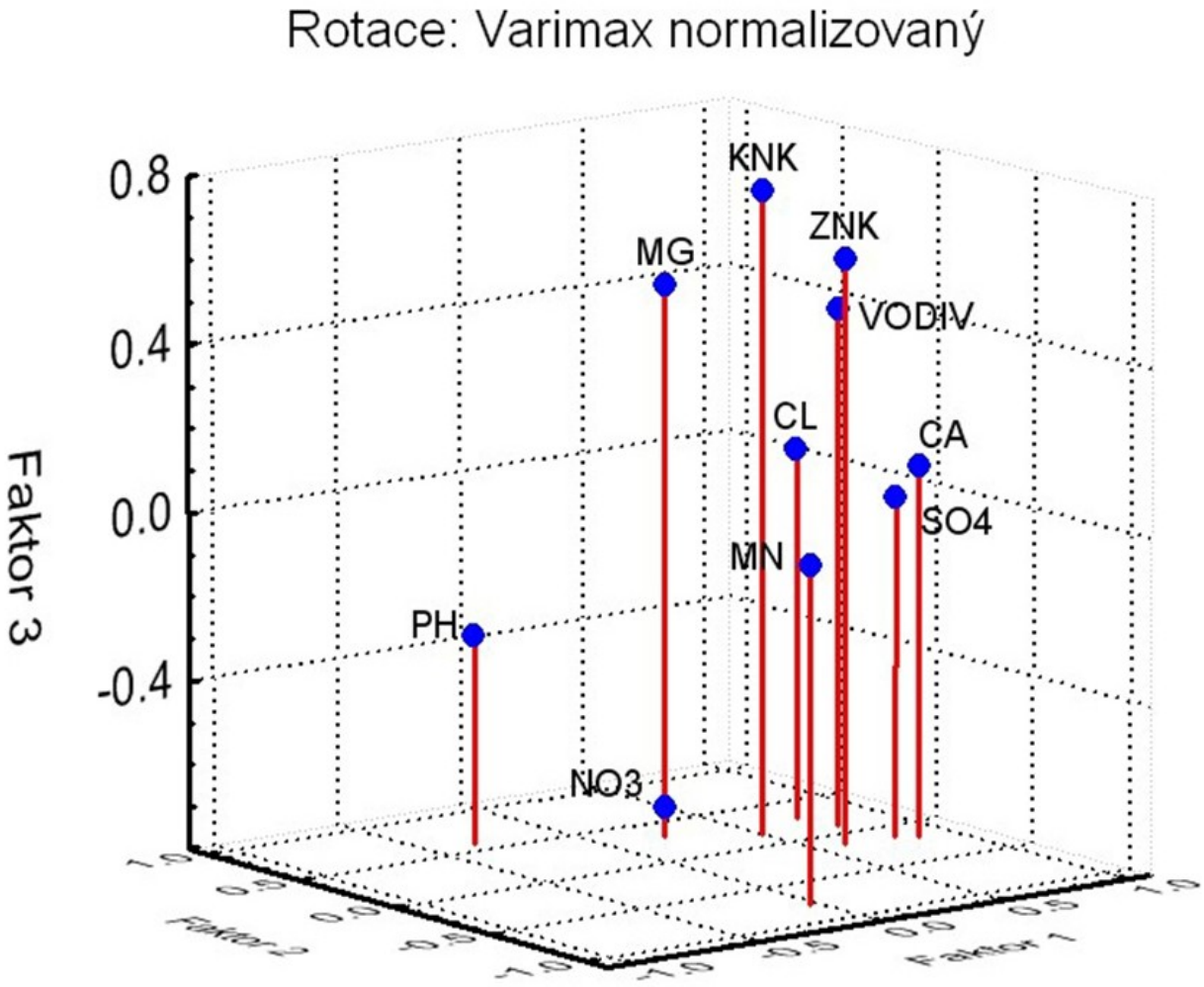

Faktor 2 : 14.94%
Faktor 1 : 42.44%
MN
ZNK
CA
KNK
SO4
VODIV
CL
MG
NO3
PH


Faktor 3 : 11.53%
Faktor 1 : 42.44%
KNK
MG
ZNK
PH
VODIV
CL
MN
CA
SO4
NO3


Faktor 3 : 11.53%
Faktor 2 : 14.94%
MG
KNK
PH
ZNK
VODIV
CL
MN
SO4
CA
NO3


Rotace: Bez rotace
Faktor 2
Faktor 1
1.0
0.5
0.0
-0.5
-1.0
MN
ZNK
CA
KNK
SO4
VODIV
MG
CL
NO3
PH


Rotace: Varimax normalizovaný
Faktor 2
Faktor 1
1.0
0.5
0.0
-0.5
-1.0
PH
MG
NO3
CL
KNK
VODIV
SO4
ZNK
CA
MN


Rotace: Varimax normalizovaný
Faktor 3
0.8
0.4
0.0
-0.4
KNK
ZNK
MG
VODIV
CL
CA
SO4
MN
PH
NO3
Faktor 2
Faktor 1

Korelační matice z původní matice všech 16 znaků

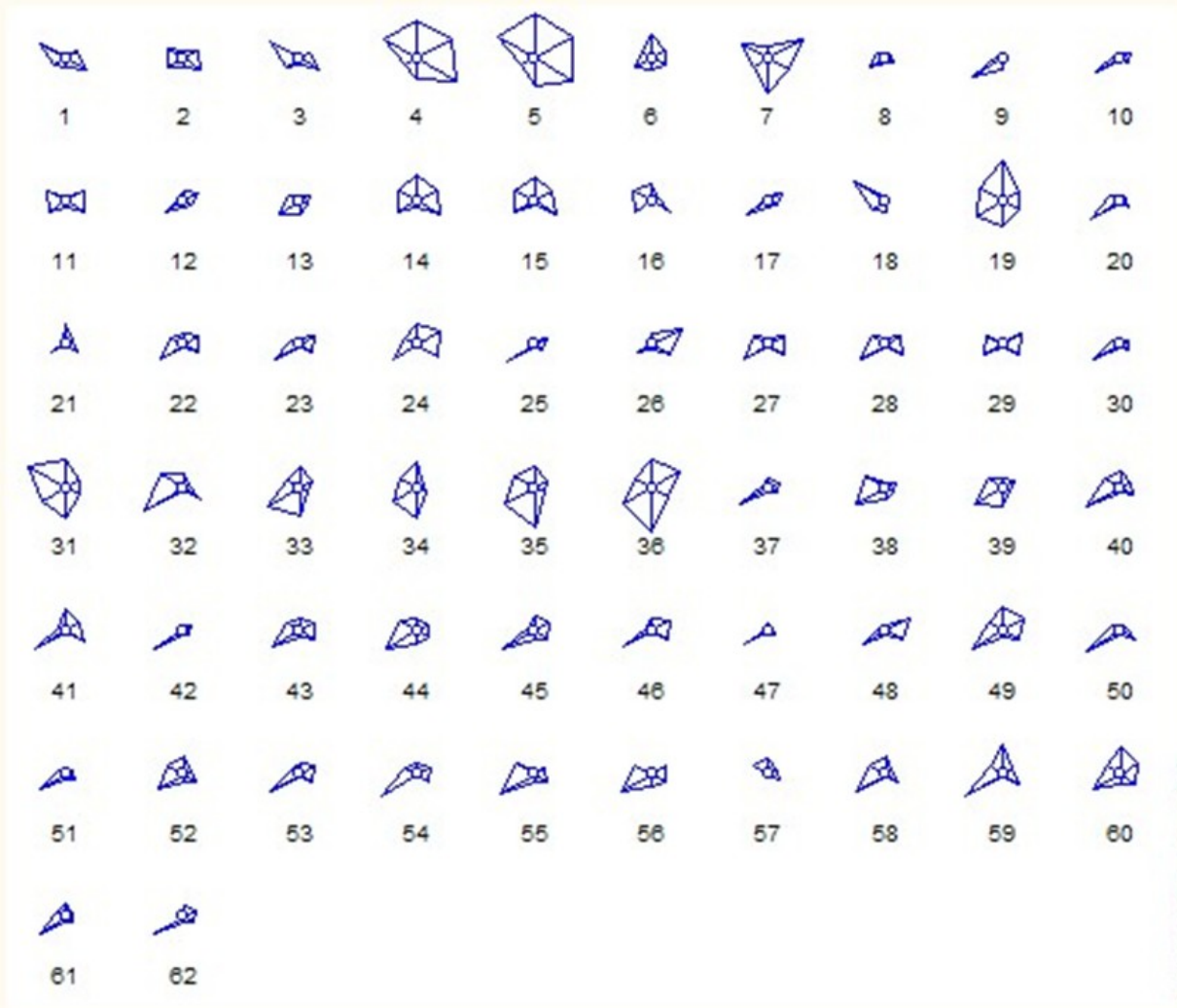
	Korelace (E404) Označ. korelace jsou významné na hlad. p < .05000 N=62 (Celé případy vynechány u ChD)													
Proměnná	Průměry	Sm.odch.	NO3	NO2	CL	CL2	SO4	PO4	NH4	CA	MG	FE	MN	PH
NO3	14.6016	15.3369	1.000000	0.047190	0.075647	0.081276	-0.053931	-0.020737	-0.109845	-0.017772	-0.123426	-0.085330	-0.215441	0.157820
NO2	0.0658	0.1387	0.047190	1.000000	0.040778	0.042804	-0.126076	*0.465180*	-0.091387	-0.175622	-0.186858	0.049748	-0.042447	*-0.358610*
CL	29.1452	21.1015	0.075647	0.040778	1.000000	*0.998609*	*0.456812*	0.032834	-0.137223	*0.539559*	*0.476512*	0.012382	-0.048272	-0.062089
CL2	29.0026	21.2213	0.081276	0.042804	*0.998609*	1.000000	*0.466434*	0.033436	-0.139016	*0.550111*	*0.473684*	0.013248	-0.047532	-0.051200
SO4	116.8598	61.3146	-0.053931	-0.126076	*0.456812*	*0.466434*	1.000000	-0.197464	0.050398	*0.755038*	*0.381889*	*0.316056*	0.042870	*-0.283529*
PO4	0.1636	0.5953	-0.020737	*0.465180*	0.032834	0.033436	-0.197464	1.000000	-0.055856	-0.075953	-0.138298	0.028380	-0.013695	-0.056366
NH4	0.1013	0.2626	-0.109845	-0.091387	-0.137223	-0.139016	0.050398	-0.055856	1.000000	0.060288	-0.082867	-0.075350	*0.399794*	-0.215715
CA	131.5323	48.9783	-0.017772	-0.175622	*0.539559*	*0.550111*	*0.755038*	-0.075953	0.060288	1.000000	0.155819	0.082685	0.028817	*-0.285059*
MG	18.6290	18.3600	-0.123426	-0.186858	*0.476512*	*0.473684*	*0.381889*	-0.138298	-0.082867	0.155819	1.000000	*0.369682*	-0.157013	-0.019601
FE	0.0560	0.1310	-0.085330	0.049748	0.012382	0.013248	*0.316056*	0.028380	-0.075350	0.082685	*0.369682*	1.000000	-0.048317	-0.106596
MN	0.2111	0.6011	-0.215441	-0.042447	-0.048272	-0.047532	0.042870	-0.013695	*0.399794*	0.028817	-0.157013	-0.048317	1.000000	-0.249614
PH	7.5281	0.2644	0.157820	*-0.358610*	-0.062089	-0.051200	*-0.283529*	-0.056366	-0.215715	*-0.285059*	-0.019601	-0.106596	-0.249614	1.000000
KNK	5.5919	1.8863	-0.232680	-0.157834	*0.406614*	*0.389809*	*0.265703*	0.028287	0.094545	*0.515586*	*0.495420*	0.173567	-0.126195	*-0.297318*
ZNK	1.4339	0.9006	-0.226557	-0.067831	*0.344606*	*0.331544*	*0.441636*	0.205221	0.091070	*0.654338*	*0.310340*	0.143989	0.004198	*-0.411754*
VODIV	684.0000	270.7810	-0.050967	-0.106941	*0.648153*	*0.621040*	*0.515015*	-0.046916	0.027953	*0.609324*	*0.449503*	0.185218	-0.063259	*-0.342510*
NEROZP	0.1268	0.1127	-0.047249	-0.055412	0.072002	0.067766	-0.054792	-0.126639	0.005071	0.072844	0.021429	0.137923	-0.050188	-0.090398

	Korelace (E404) Označ. korelace jsou významné na hlad. p < .05000 N=62 (Celé případy vynechány u ChD)			
Proměnná	KNK	ZNK	VODIV	NEROZP
NO3	-0.232680	-0.226557	-0.050967	-0.047249
NO2	-0.157834	-0.067831	-0.106941	-0.055412
CL	*0.406614*	*0.344606*	*0.648153*	0.072002
CL2	*0.389809*	*0.331544*	*0.621040*	0.067766
SO4	*0.265703*	*0.441636*	*0.515015*	-0.054792
PO4	0.028287	0.205221	-0.046916	-0.126639
NH4	0.094545	0.091070	0.027953	0.005071
CA	*0.515586*	*0.654338*	*0.609324*	0.072844
MG	*0.495420*	*0.310340*	*0.449503*	0.021429
FE	0.173567	0.143989	0.185218	0.137923
MN	-0.126195	0.004198	-0.063259	-0.050188
PH	*-0.297318*	*-0.411754*	*-0.342510*	-0.090398
KNK	1.000000	*0.759789*	*0.733499*	0.219339
ZNK	*0.759789*	1.000000	*0.693455*	-0.103607
VODIV	*0.733499*	*0.693455*	1.000000	*0.253084*
NEROZP	0.219339	-0.103607	*0.253084*	1.000000

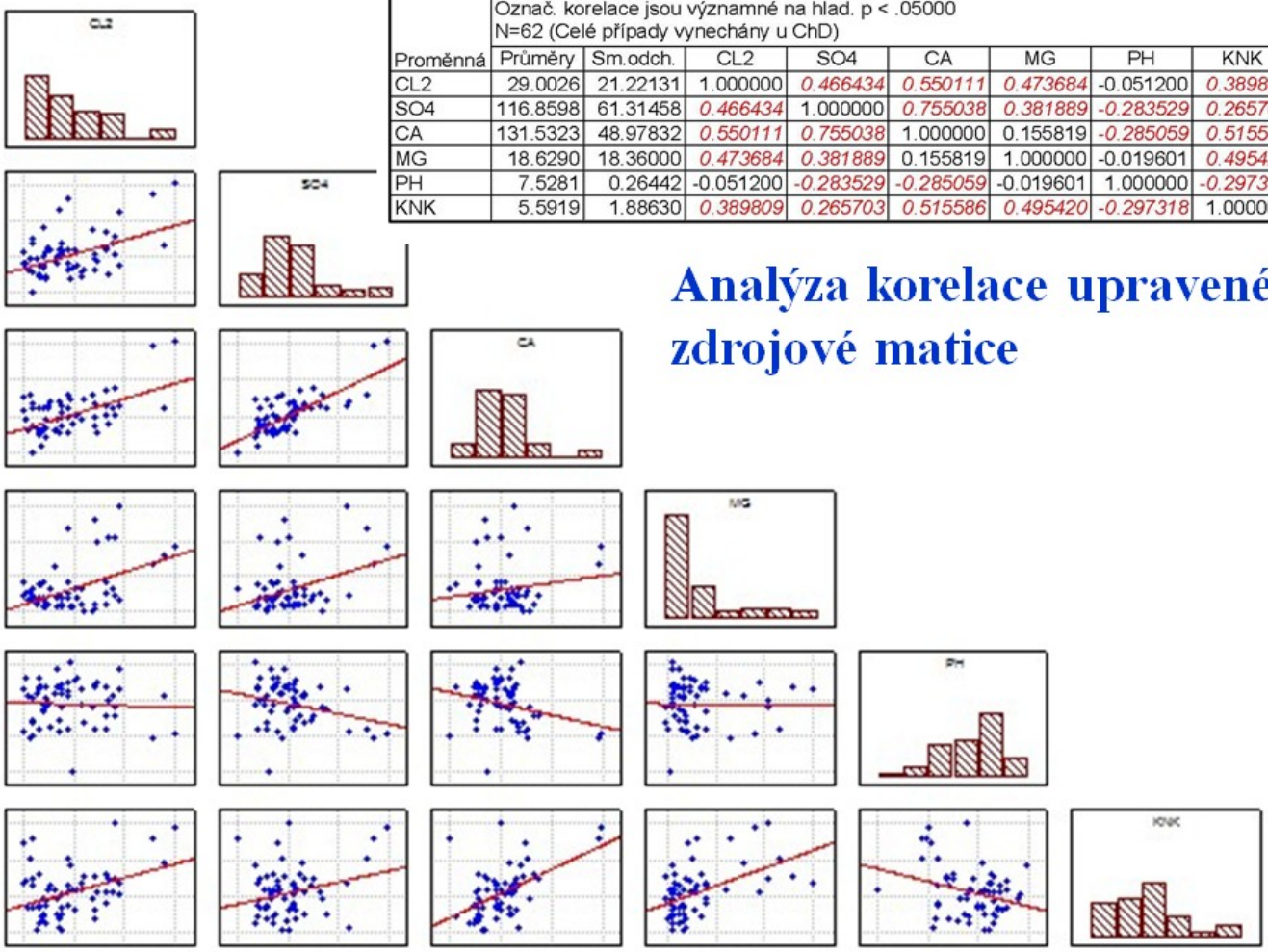
Korelační matice u výběru znaků v největší korelaci:

CL2, SO4, CA, MG, PH, KNK

	Korelace (E404) Označ. korelace jsou významné na hlad. p < .05000 N=62 (Celé případy vynechány u ChD)							
Proměnná	Průměry	Sm.odch.	CL2	SO4	CA	MG	PH	KNK
CL2	29.0026	21.22131	1.000000	*0.466434*	*0.550111*	*0.473684*	-0.051200	*0.389809*
SO4	116.8598	61.31458	*0.466434*	1.000000	*0.755038*	*0.381889*	*-0.283529*	*0.265703*
CA	131.5323	48.97832	*0.550111*	*0.755038*	1.000000	0.155819	*-0.285059*	*0.515586*
MG	18.6290	18.36000	*0.473684*	*0.381889*	0.155819	1.000000	-0.019601	*0.495420*
PH	7.5281	0.26442	-0.051200	*-0.283529*	*-0.285059*	-0.019601	1.000000	*-0.297318*
KNK	5.5919	1.88630	*0.389809*	*0.265703*	*0.515586*	*0.495420*	*-0.297318*	1.000000

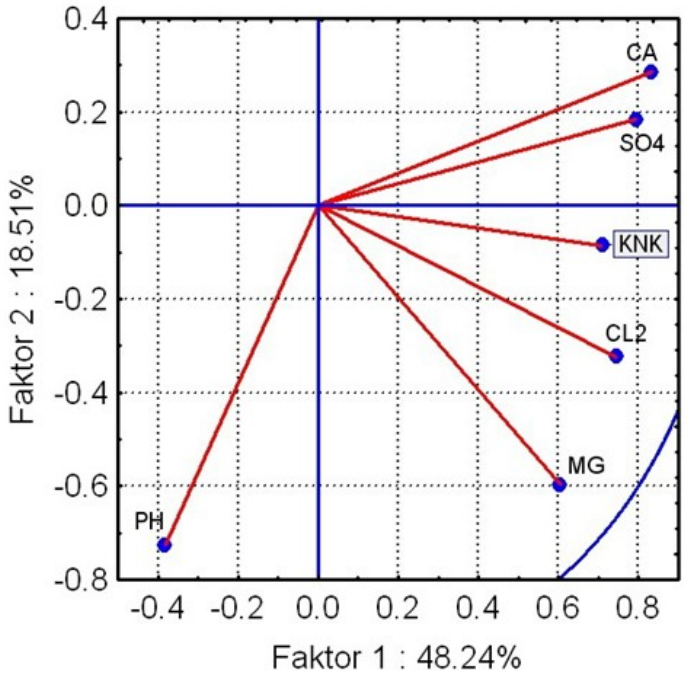


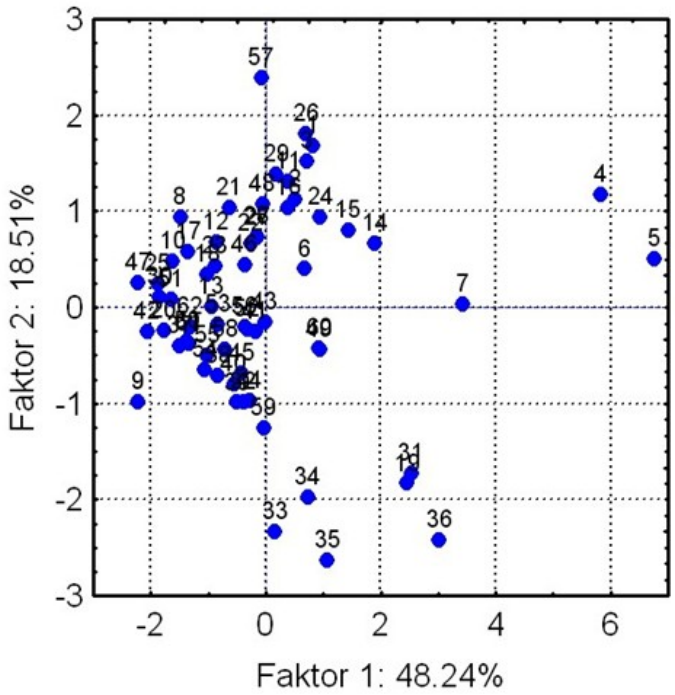
Proměnná	Korelace (E404) Označ. korelace jsou významné na hlad. p < .05000 N=62 (Celé případy vynechány u ChD) Průměry	Sm.odch.	CL2	SO4	CA	MG	PH	KNK
CL2	29.0026	21.22131	1.000000	*0.466434*	*0.550111*	*0.473684*	-0.051200	*0.389809*
SO4	116.8598	61.31458	*0.466434*	1.000000	*0.755038*	*0.381889*	*-0.283529*	*0.265703*
CA	131.5323	48.97832	*0.550111*	*0.755038*	1.000000	0.155819	*-0.285059*	*0.515586*
MG	18.6290	18.36000	*0.473684*	*0.381889*	0.155819	1.000000	-0.019601	*0.495420*
PH	7.5281	0.26442	-0.051200	*-0.283529*	*-0.285059*	-0.019601	1.000000	*-0.297318*
KNK	5.5919	1.88630	*0.389809*	*0.265703*	*0.515586*	*0.495420*	*-0.297318*	1.000000



Analýza korelace upravené zdrojové matice

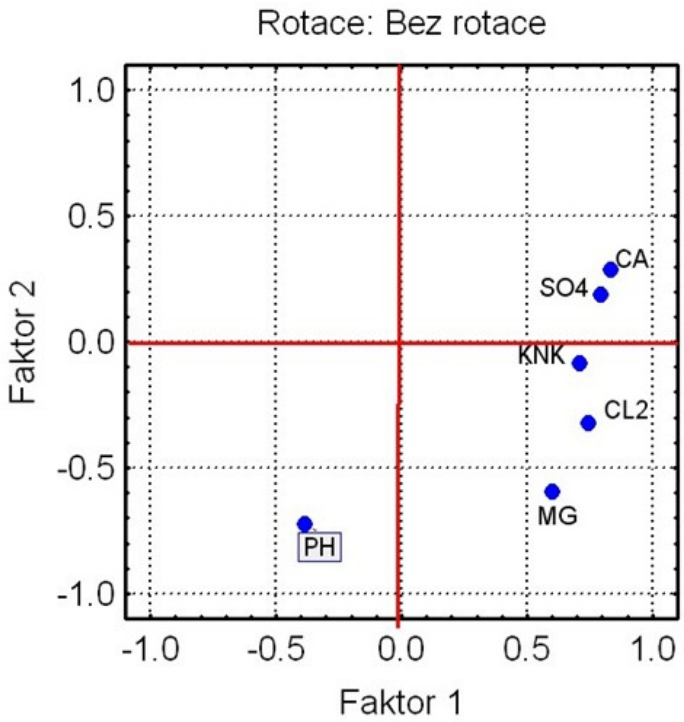
PCA zdrojové matice po výběru znaků v největší korelaci:
CL2, SO4, CA, MG, PH, KNK

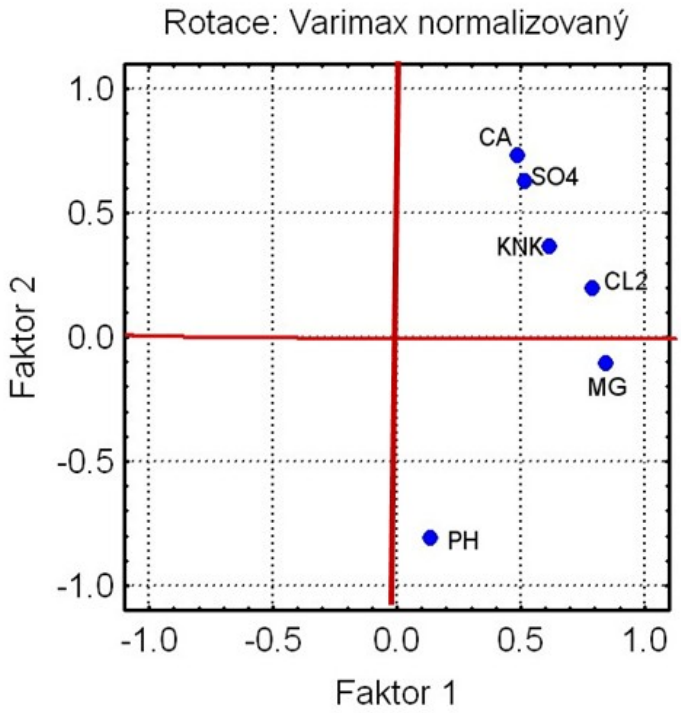


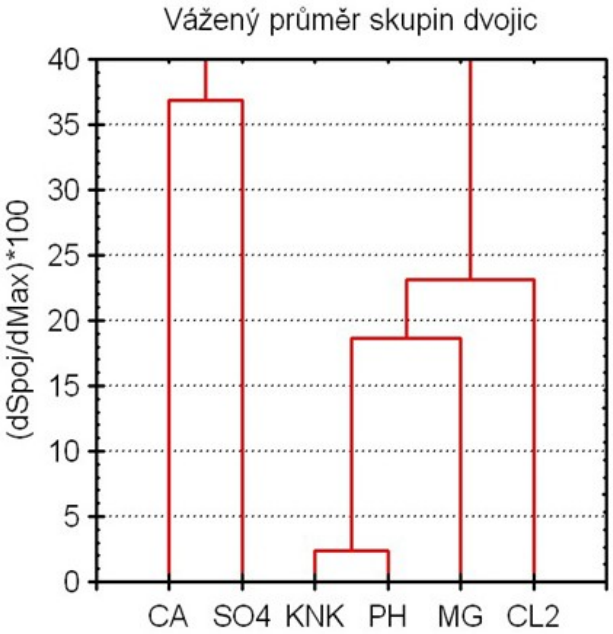


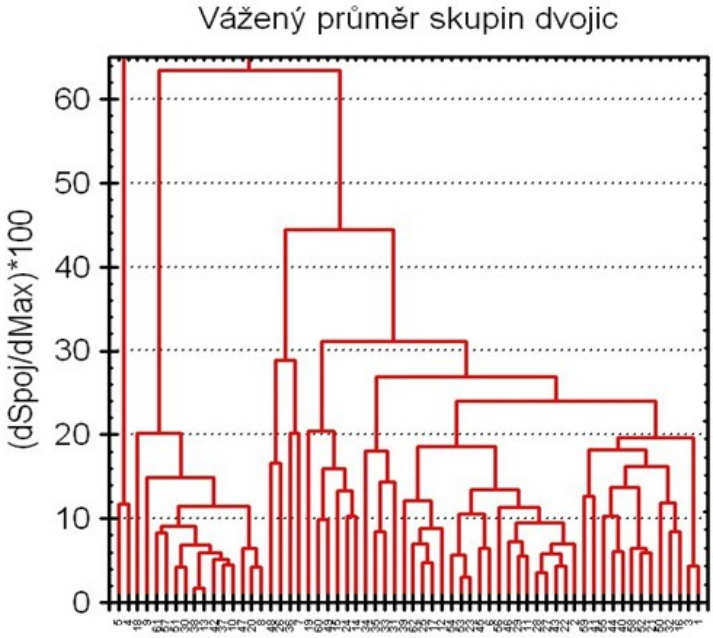
Analýza korelace vybraných šesti znaků, které nejvíce ovlivňují kvalitu pitné vody.

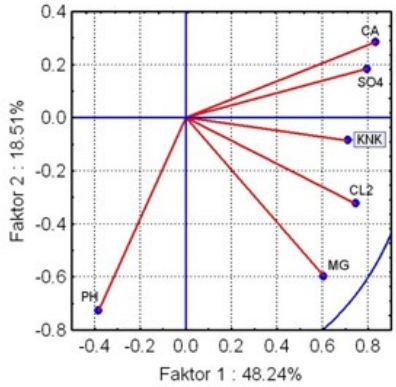
Snadnější monitorování outlierů kvality pitné vody v šesti-rozměrném prostoru pouze vybraných znaků.

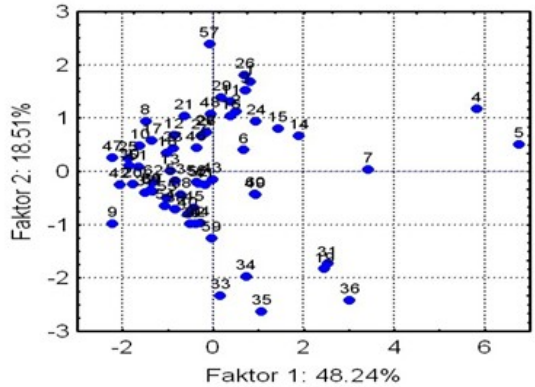












Závěr:

1. Bylo nalezeno 6 nejdůležitějších znaků, dle kterých lze objektivně posoudit kvalitu 62 vzorků pitné vody(Korelace, PCA).

2. V 62 vzorcích pitné vody byly vytvořeny shluky vody podobných šesti sledovaných vlastností (EDA, PCA, CLU).

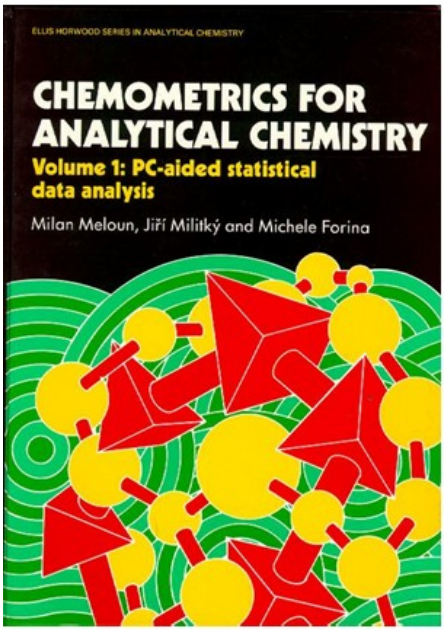
3. Byly nalezeny odlehlé vzorky, které nepatří do žádného shluku (EDA, PCA, CLU).

4. Klasifikace vícerozměrnou statistickou analýzou dat byla porovnána s chemickými závěry o vzorcích pitné vody a byl shledán souhlas.

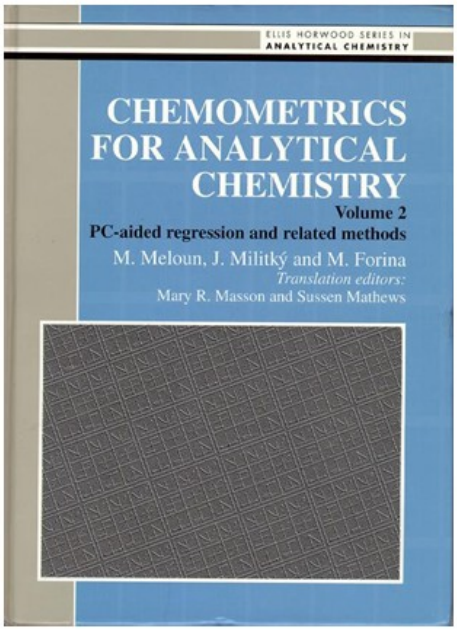
Statistické zpracování experimentálních dat

Naše učebnice chemometrie v angličtině

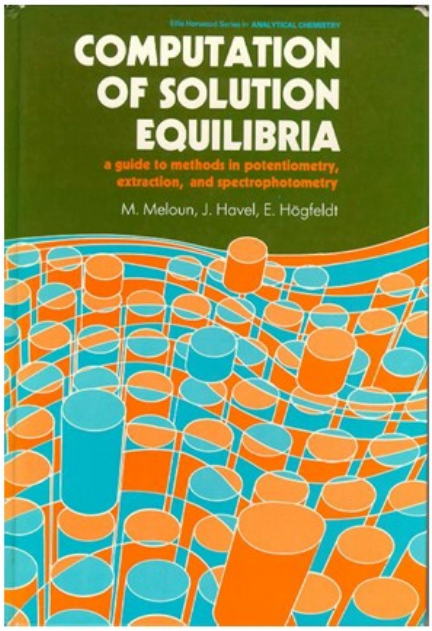
(Ellis Horwood, Chichester – John Wiley & Son, New York)



1992



1994

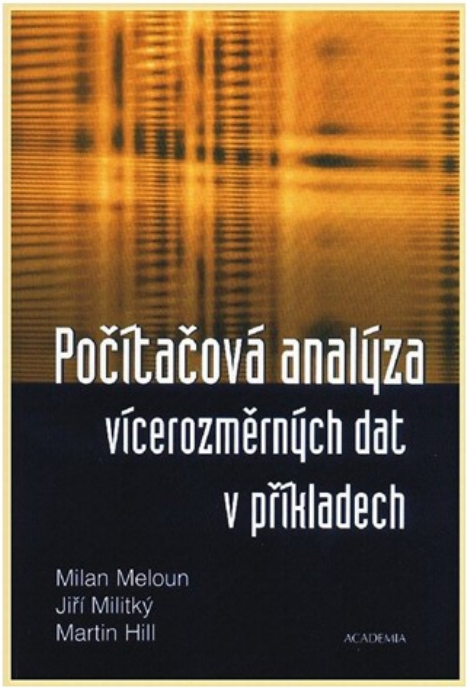


1986

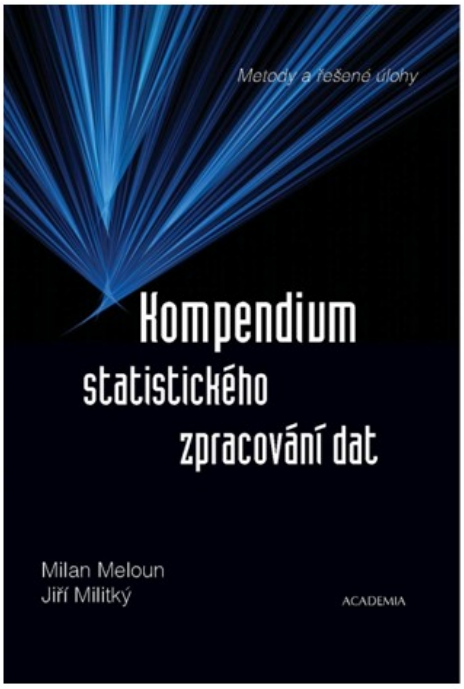
Statistické zpracování experimentálních dat

Potřebná literatura ke studiu v roce 2010

2004



2005



2006