Klasifikace podzemních vod diskriminační analýzou

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice,
532 10 Pardubice, milan.meloun@upce.cz,

a Jindřich Freisleben
Český hydrometeorologický ústav, Na Šabatce 17,
143 06 Praha 4 – Komořany, freisleben@chmi.cz,

Souhrn: *Diskriminační analýza umožňuje hodnocení rozdílů mezi dvěma nebo více skupinami objektů charakterizovaných více znaky-diskriminátory. Obyčejně se dále dělí na techniky, které interpretují rozdíly mezi předem stanovenými skupinami objektů, a techniky, kde je cílem klasifikace objektů do skupin. Jsou porovnávány diskriminátory každého objektu (například charakteristiky sloučenin, vlastnosti objektu, vzorku vody, atd.) se znaky ostatních objektů. Na základě podobností nebo rozdílů se pak provede klasifikace vzorků podzemních vod buď čistě subjektivně na základě zkušeností, nebo objektivními metodami. Na základě diagramů diskriminačního skóre jsou klasifikovány vzorky podzemních vod do tří tříd. Diagramy poskytují vizuální ověření, jak diskriminační funkce zařazují objekty do tříd. Všechny diskriminátory naměřené chemickou analýzou nebyly shledány vhodné pro dostatečně přesné přiřazení objektů podzemních vod lineární diskriminační analýzou. Procento správně zařazených objektů v rámci třídy mělkých vrtů je dost nízké, pouze 58 %. Příčina může být jednak v tom, že monitorované ukazatele nemají dostatečnou diskriminační „sílu" a také v tom, že většina diskriminátorů vykazuje jiné než normální rozdělení.*

Klíčová slova: *DA, PCA, CM, Cattelův graf, diskriminace, klasifikace, diskriminační skóre, Fisherova diskriminační funkce*

1. Úvod

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do existující třídy (*diskriminační analýza DA*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, například, když rozdělení náhodného vektoru charakterizujícího objekty je normální, pak hovoříme o *parametrických klasifikačních metodách.* Není-li klasifikace založena na znalostech rozdělení náhodného vektoru, mluvíme o *neparametrických klasifikačních metodách.* Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také *vícerozměrné škálování MDS.*

2. Podstata metody DA

Klasická klasifikační diskriminační analýza, zavedená Ronaldem Fisherem v roce 1936, patří mezi metody zkoumání vztahu mezi skupinou *p* nezávislých znaků, zvaných *diskriminátory* (sloupců zdrojové matice), a jednou kvalitativní závisle proměnnou – výstupem. Výstupem je v nejjednodušším případě binární proměnná *y*, nabývající hodnotu 0 pro případ, že objekt je v první třídě, respektive hodnotu 1 pro případ, že objekt je ve druhé třídě. O třídách je známé, že jsou zřetelně odlišené a každý objekt patří do jedné z nich. Účelem může být také identifikace, které znaky přispívají do procesu klasifikace. Ve vstupních datech trénovací skupiny jsou svými hodnotami diskriminátorů a výstupů v š e c h n y objekty zařazené do tříd. Účelem je nalézt predikční model umožňující zařadit nové objekty do tříd.

2.1 Zařazovací pravidla DA

Pro zjednodušení uvažujme, že účelem je klasifikace do jedné ze dvou tříd (A, B) a že klasifikace se provádí na základě jednoho znaku *x* s normálním rozdělením. Ve třídě A jde o rozdělení $N(\mu_A, \sigma_A^2)$ a ve třídě B jde o rozdělení $N(\mu_B, \sigma_B^2)$. Nový objekt nechť má hodnotu *x*. Je logické vybrat tu třídu, pro kterou je *x* blíže ke střední hodnotě dané třídy. Můžeme tedy určit *prahový bod* $C = (\mu_A + \mu_B)/2$. Pro případy, kdy $x < C$ se pak objekt zařadí do třídy (A) a pro $x \geq C$ do druhé třídy (B). Tato pravděpodobnost je pro obě kategorie stejná. Pokud by se toto pravidlo použilo i pro případ nestejných rozptylů obou rozdělení, kdy například $\sigma_A^2 < \sigma_B^2$, došlo by k situaci, že pravděpodobnost nesprávné klasifikace pro třídu A by vyšla větší než pro třídu B. Bude tedy třeba penalizovat *C* s ohledem na nestejný rozptyl. Je zřejmá analogie s *t*-testem porovnání dvou středních hodnot pro nestejné rozptyly. Pro případ dvou diskriminátorů (x_1 x_2) je výhodné zobrazovat v prostoru x_1 x_2 elipsy konstantní hustoty, například pro pravděpodobnost 0.95. Záleží na tom, zda jsou kovarianční matice C_A a C_B shodné či nikoliv.

2.2 Lineární (LDA) a kvadratická (QDA) diskriminační funkce

Při formálním odvození tvaru diskriminačních funkcí je možné vyjít z rovnice pro aposteriorní pravděpodobnost příslušnosti k *j*-té skupině ($j = 1, 2$) a hledat její

maximum. Podle typu hustot pravděpodobnosti pro znaky $f_1(\boldsymbol{x})$ a $f_2(\boldsymbol{x})$ se tak liší jednotlivé diskriminační metody v tom, jak jsou specifikovány dělicí oblasti a jaký mají tvar: pro normální rozdělení, lišící se jen středními hodnotami tříd, dostáváme lineární diskriminační analýzu (LDA), pro normální rozdělení, lišící se jak středními hodnotami, tak i kovariančními maticemi tříd dostáváme kvadratickou diskriminační analýzu (QDA), pro případ směsí normálních rozdělení vycházejí nelineární diskriminační funkce, pro případ neparametrických hustot rozdělení znaků ve třídách obdržíme flexibilní diskriminační funkce, naivní Bayesův přístup spočívá v představě, že každá hustota pravděpodobnosti ve třídě je získána jako součin marginálních hustot znaků (znaky jsou zde považovány za podmíněně nezávislé). Pro případ vícerozměrného Gaussova rozdělení v *i*-té třídě má odpovídající hustota pravděpodobnosti tvar

$$f_i(x) = \frac{1}{(2\pi)^{m/2}\ \det(C_i)^{1/2}}\ \exp\left(-\frac{1}{2}\ (x - \mu_i)^{\mathrm{T}}\ C_i^{-1}\ (x - \mu_i)\right),$$

kde $m = 2$ je počet znaků.

2.3 Úprava prahového bodu

Dosud byl prahový bod *C* užíván jako bod jenž udává stejné procento chyb obojího typu, čili pravděpodobnost chybného zařazení objektu ze třídy 1 do třídy 2 a naopak. Volba prahového bodu *C* však může být provedena tak, že bude poskytovat požadovaný poměr apriorních pravděpodobností π_1 a π_2. Pro vícerozměrný normální model bude optimální volba prahového bodu *C* daná vzorcem

$$C = \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\frac{\pi_1}{\pi_2}.$$

Když bude $\pi_1 = \pi_2 = 0.5$, bude *C* rovno jednoduššímu výrazu bez ln. Stejně jako v regresní analýze nejsou hodnoty koeficientů $a_1, a_2, ..., a_p$ přímo porovnatelné. Relativní vliv na každou proměnnou v diskriminační funkci můžeme získat ze *standardizovaných diskriminačních koeficientů.* Tato technika se týká užití společné kovarianční matice. Standardizované koeficienty se vypočtou vynásobením koeficientů a_i odpovídající směrodatnou odchylkou s_i.

2.4 Volba diskriminátorů

Otázkou je, zda volba diskriminátorů x, je schopna zajistit dostatečně přesné zařazení objektů do tříd, tj. diskriminaci. Byla navržena řada postupů jak provést selekci znaků. Principem většiny metod je zajištění dostatečné separability tříd a volba takových diskriminátorů, které vedou k maximalizaci nějaké míry. Jindy se volí postup, který začne se všemi původními diskriminátory a postupně se vypouštějí takové, které vedou k nedostatečné separaci. V mnoha situacích je diskriminační analýza, stejně jako lineární regresní analýza, použita jako exploratorní pomůcka. Pro nalezení vhodného modelu, je do dat zahrnuta celá paleta potenciálně využitelných znaků. Není však předem známo, které diskriminátory jsou pro zařazení objektů do tříd účinné. Jedním z možných výsledků diskriminační analýzy je také identifikace „účinných“ diskriminátorů. Při výběru „účinných“ diskriminátorů jde o analogii s vícenásobnou regresní analýzou. V diskriminační analýze se místo testování, zda se hodnota čtverce vícenásobného korelačního koeficientu R^2 změní přidáním nebo odebráním proměnné testuje, zda se změní hodnota Mahalanobisovy vzdálenosti D_M^2. Obvykle se užívají popsaná testační kritéria tak, že pro testační kritérium F se, pokud přidáváme proměnné, za α dosazuje hodnota 0.15. Bohužel nebývá známá vhodná hodnota pro α při použití F-testu, odebíráme-li proměnné, a proto je v literatuře obvykle doporučována vysoká hodnota $\alpha = 0.30$. Všechna kritéria výběru nezávisle proměnných v lineární regresní analýze platí i v diskriminační analýze. Všeobecně užívaným algoritmem je *krokový výběr diskriminátorů*, jehož principy jsou známé z lineární regrese. Postup kombinuje jak přidávání diskriminátorů, tak i jejich odstraňování. V krokové metodě má první diskriminátor, zahrnutý do modelu ve výběrovém kritériu, největší přijatelnou hodnotu. Po zavedení prvního diskriminátoru je hodnota kritéria přepočítána pro všechny diskriminátory v modelu a diskriminátor s největší přijatelnou hodnotou *zaváděcího kritéria* je zaveden do modelu jako další. V tomto okamžiku je diskriminátor, který byl zaveden do modelu jako první, znovu přepočten, zda splňuje také *odstraňovací kritérium*. Jestliže ano, je z modelu odstraněn. Dalším krokem je vyšetření diskriminátorů připravených k zavedení do modelu, následované vyšetřením diskriminátorů připravených v modelu k odstranění. Vybírání diskriminátorů se ukončí, když žádné další diskriminátory nesplňují zaváděcí nebo odstraňovací kritérium.

2.5 Kritéria pro vybírání diskriminátorů

Existuje několik rozhodovacích kritérií k vybírání diskriminátorů. U *Wilkova kritéria* λ platí že, když diskriminátor v diskriminační funkci poskytuje nejmenší hodnotu Wilkova kritéria λ, je tento diskriminátor zahrnut do modelu. K zavedení nebo odstranění diskriminátoru je dovolen jeden krok. Maximální počet kroků k vybírání diskriminátorů je roven dvojnásobku jejich počtu. Podobně jako ve vícenásobné regresi, když jsou některé nezávisle proměnné lineárními kombinacemi ostatních nezávisle proměnných, není možné očekávat jediné řešení. Aby se předešlo výpočetním problémům, je před zavedením diskriminátoru do modelu stanovena jeho *tolerance.* Tolerance je mírou lineární asociace mezi diskriminátory a vypočte se pro *i*-tý diskriminátor dle $1 - R_i^2$, kde R_i^2 je čtverec vícenásobného korelačního koeficientu, když je uvažován *i*-tý diskriminátor za závisle proměnnou a když je uvažována regresní rovnice mezi tímto *i*-tým diskriminátorem a ostatními diskriminátory. Malé hodnoty tolerance indikují, že *i*-tý diskriminátor je tvořen lineární kombinací ostatních diskriminátorů. Diskriminátory s tolerancí menší než 0.001 není však vhodné do modelu zařadit.

Významnost změny Wilkova kritéria λ po zavedení diskriminátoru do modelu nebo odstranění z modelu je založena na testačním kritériu *F*. Aktuální hodnota testačního kritéria *F* nebo vypočtená statistická významnost α slouží jako kritérium pro zavedení diskriminátoru do modelu nebo k jeho odstranění. Obě kritéria však nemusí být ekvivalentní, protože pevné hodnoty kvantilu *F* mají rozdílnou pravděpodobnost v závislosti na počtu diskriminátorů v modelu. Aktuální vypočtená statistická významnost, spojená s kvantilem *F* při zavedení a s kvantilem *F* při odstranění, není obyčejně vypočtena z rozdělení *F*, protože je zde vyšetřeno mnoho diskriminátorů a jsou vybrány největší a nejmenší hodnota *F*. Skutečnou hladinu významnosti α je obtížné vyčíslit, protože závisí na mnoha faktorech včetně uvažované korelace mezi diskriminátory.

Dříve než začne pracovat krokový algoritmus, jsou na začátku v nultém kroku jak tolerance, tak i minimum tolerance položeny rovné 1, protože v modelu dosud nejsou diskriminátory. Vedle Wilkova kritéria λ se pro statistickou významnost každého diskriminátoru vyčísluje také *F*-test. Hodnota *F* pro změnu Wilkova kritéria λ při

přidání diskriminátoru do modelu tak, že model obsahuje celkem *p* diskriminátorů, se vyčíslí dle vztahu

$$F_{\text{změny}} = \frac{n - g - p}{g - 1} \left(\frac{\frac{1 - \lambda_{p+1}}{\lambda_p}}{\frac{\lambda_{p+1}}{\lambda_p}} \right),$$

kde *n* je celkový počet objektů, *g* udává počet tříd, λ_p značí Wilkovo lambda před přidáním diskriminátoru a λ_{p+1} je Wilkovo lambda po přidání diskriminátoru do modelu. V každém kroku je ten diskriminátor, který způsobuje nejmenší hodnotu Wilkova kritéria λ, zařazen do modelu. Vedle Wilkova kritéria λ existují ještě další kritéra.

Mahalanobisova vzdálenost $D_{1,2}^2$ je zobecněná míra vzdálenosti mezi dvěma třídami 1 a 2 a je definována vztahem

$$D_{1,2}^2 = (n - g) \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} (\bar{x}_{i1} - \bar{x}_{i2})(\bar{x}_{j1} - \bar{x}_{j2}),$$

kde *m* udává počet diskriminátorů v modelu, $\bar{x}_{i1}$ je průměr *i*-tého diskriminátoru ve třídě 1. Protože je Mahalanobisova vzdálenost $D_{1,2}^2$ kritériem pro volbu diskriminátorů, je toto kritérium všech párů tříd vyčísleno jako první. Ten diskriminátor, který měl největší hodnotu D_M^2 pro dvě od začátku nejtěsnější třídy, čili které měly nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu.

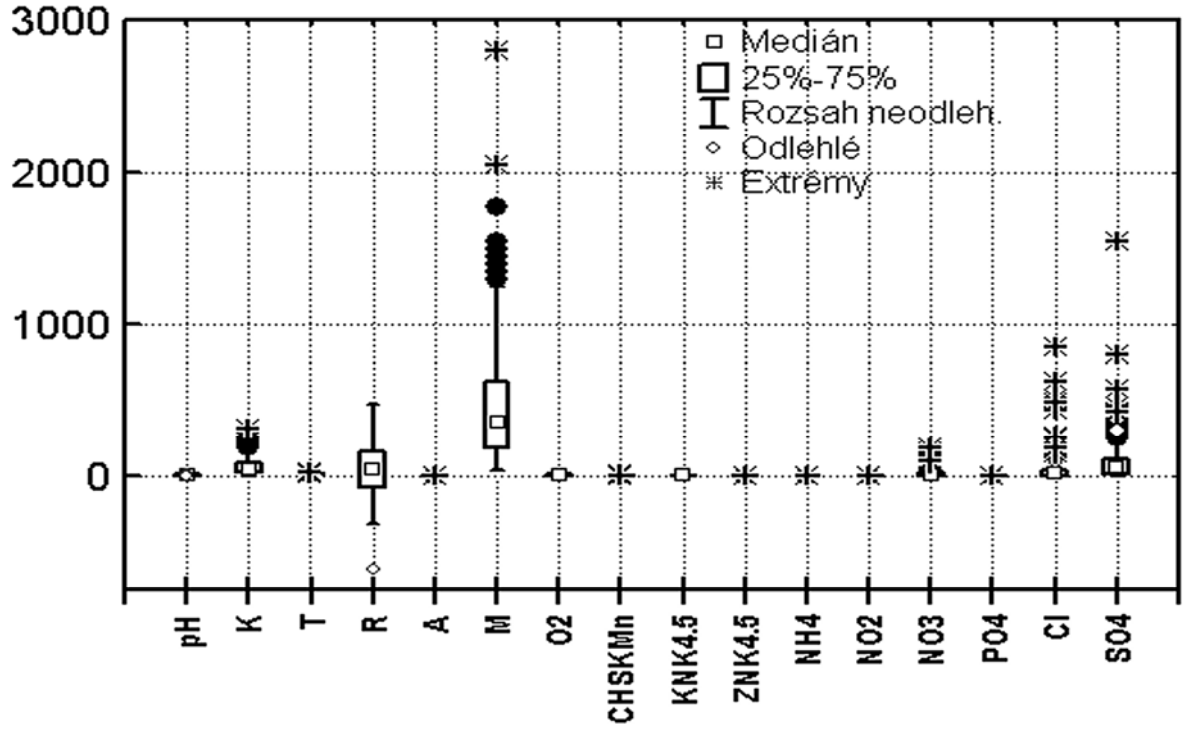
3. Výsledky: *Monitoring jakosti podzemních vod*

V rámci monitoringu jakosti podzemních vod bylo analyzováno 16 diskriminátorů ve 462 vzorcích podzemních vod. Byly navrženy tři základní kategorie sledovaných vod, a to vody z pramenů, vody z mělkých vrtů a konečně vody z hlubokých vrtů. Diskriminační analýzou (DA) je nyní třeba zjistit u dosud nezařazených výsledků vod dle jejich naměřených kvantitativních obsahů sloučenin či vlastností, představujících zde diskriminátory nejpravděpodobnější typ souboru podzemních vod, ze kterých byly vzorky vod odebrány.

- **Data:** Data zdrojové matice souboru 462 řádků vzorků podzemních vod a 16 sloupců diskriminátorů jsou rozdělena na dva stejně velké výběry, a to jednak na výběr všech lichých 231 řádků vzorků vod představujících zde analyzovaný soubor k výstavbě

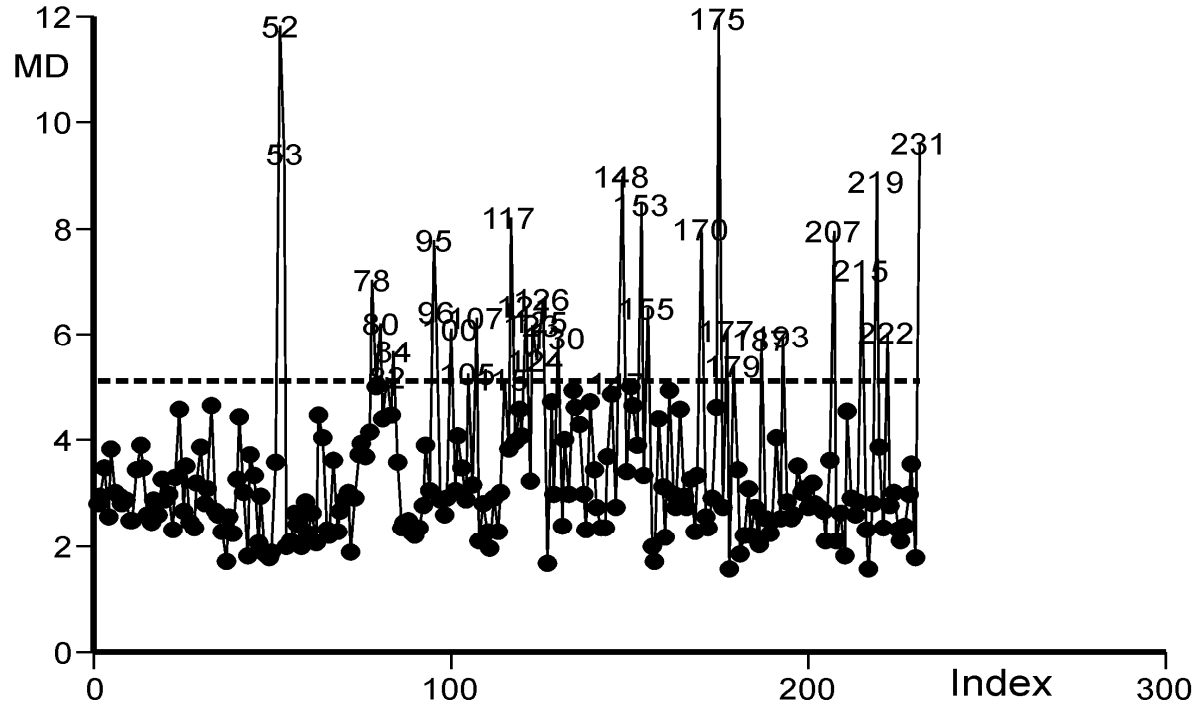
diskriminačního modelu, a jednak na výběr všech sudých 231 řádků vzorků, představující klasifikovaný soubor k vlastnímu otestování navrženého diskriminačního modelu. Za *závisle proměnnou* se budeme brát znak *TO*, který značí typ objektu, nabývající tří hodnot: 0 značí pramen, 1 značí mělký vrt, a konečně 2 značí hluboký vrt. Za *nezávisle proměnné* se bude považovat 16 diskriminátorů: *pH* značí naměřené pH vody v laboratoři, *K* je naměřená hodnota konduktivity [mS/m], *T* je teplota vody [°C], *R* je oxidačně redukční potenciál [mV], *A* absorbance změřená při vlnové délce 254nm v kyvetě délky 1cm, *M* je celková mineralizace [mg/l], O_2 obsah kyslíku rozpuštěného [mg/l], *CHSKMn* chemická spotřeba kyslíku manganistanem [mg/l], *KNK*4,5 kyselinová neutralizační kapacita do pH 4,5 [mmol/l], *ZNK*8,3 zásadová neutralizační kapacita do pH 8,3 [mmol/l], NH_4 obsah amonných iontů [mg/l], NO_2 obsah dusitanů [mg/l], NO_3 obsah dusičnanů [mg/l], PO_4 obsah fosforečnanů [mg/l], *Cl* obsah chloridů [mg/l], SO_4 obsah síranů [mg/l].

- **Řešení: (a) Exploratorní analýza diskriminátorů:** Zdrojová matice dat analyzovaného výběru obsahuje 231 řádků vzorků vod a 16 sloupců vyšetřovaných diskriminátorů a neobsahuje žádné chybějící prvky. Žádný objekt v řádku není třeba vyřadit pro nedostatečný popis diskriminátorů.



Obr. 1 Krabicové grafy všech diskriminátorů ukazují na míru rozptýlení a proměnlivost

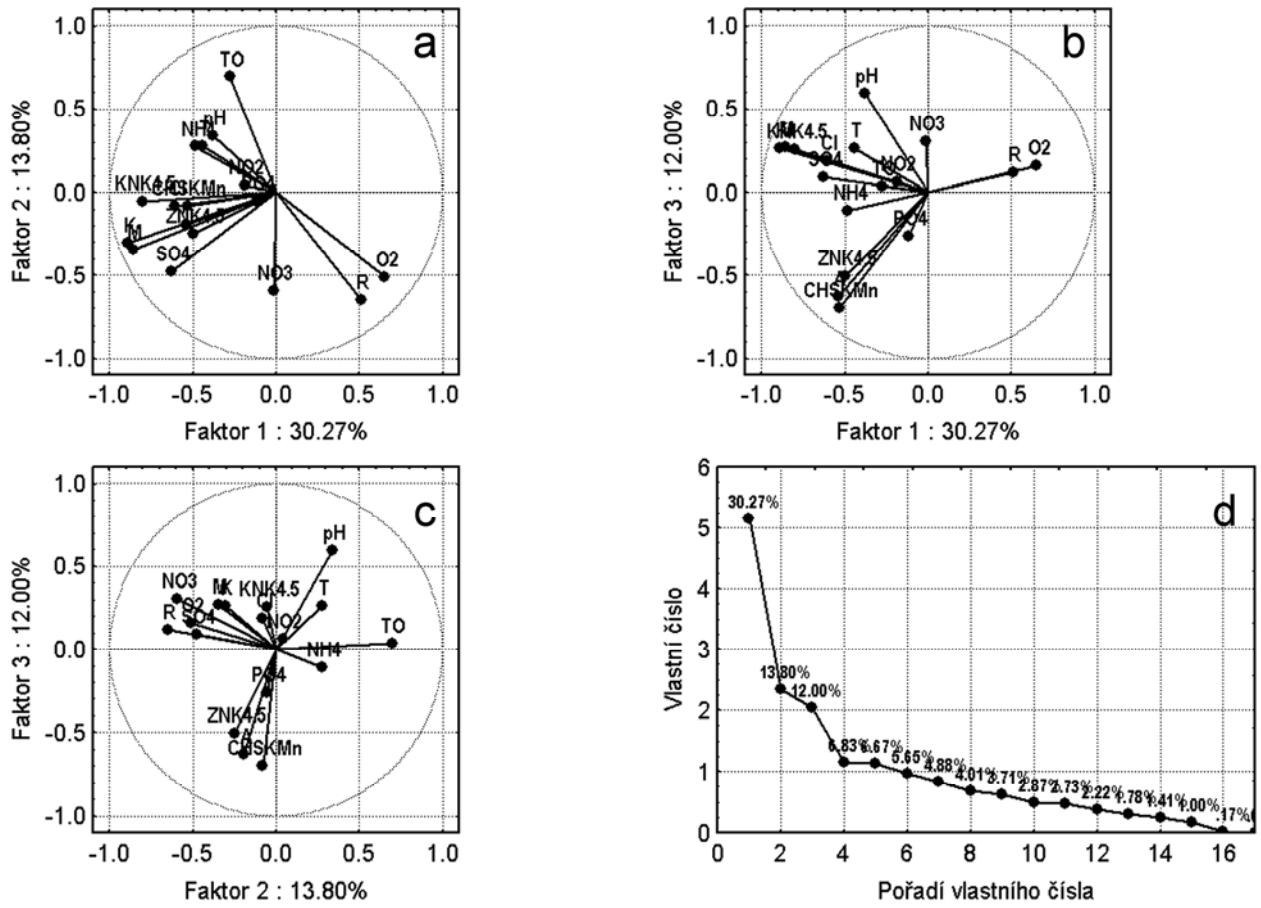
Byl použit statistický program STATISTICA 7.0 (StatSoft Praha). Nejprve se provede exploratorní analýza dat EDA a vyčíslení popisných statistik (Tabulka 1).



Obr. 2 Indexový graf Mahalanobisovy vzdálenosti objektu od těžiště indikuje odlehlé hodnoty

V rámci EDA ukazuje krabicový graf všech diskriminátorů na obr. 1 dostatečnou proměnlivost většiny znaků. Graf Mahalanobisovy vzdálenosti na obr. 2 upozorňuje na indikovanou spoustu vybočujících vzorků vod, o kterých je třeba dále věcně uvažovat, zda by neměly být z další analýzy vyřazeny či zda do výběru skutečně patří. Základním předpokladem diskriminační analýzy (DA) je platnost vícerozměrné normality v rámci tříd, a dále nutnou (ne však postačující) podmínkou je také normální rozdělení dat jednotlivých diskriminátorů. Normální rozdělení se totiž v rámci popisných charakteristik polohy, rozptýlení a tvaru rozdělení projevuje blízkými hodnotami aritmetického průměru a mediánu, dále pak šikmost rozdělení je blízká 0 a špičatost blízká 3. Z těchto hledisek není normalita splněna u všech znaků. Nejmarkantnější odchylky od normality se objevují u diskriminátorů *A, NH4, NO2, NO3, PO4, Cl* a *SO4*. Je nutné si proto uvědomit, že data by bylo vhodné zpracovat také jinou vícerozměrnou statistickou metodou, méně citlivou na normalitu rozdělení jako je například logistická regrese. Dalším předpokladem DA je podobnost kovariačních matic tříd, a tím pádem i přibližně stejně velkých směrodatných odchylek v rámci jednotlivých tříd. Větší rozdíly

jsou v míře rozptýlení shledány u diskriminátorů *K, M, Cl* a *SO4*. Nejdůležitější vlastností diskriminátoru je jeho dostatečný příspěvek k separaci objektů mezi třídami. To plyne z rozdílných hodnot průměrů jednotlivých tříd. Vzhledem k této vlastnosti lze předběžně za nevhodné diskriminátory označit znaky: *pH, T, CHSKMn, PO4*. Zdá se, že lze docílit dobré separace mezi třídou **MV** (mělké vrty) a ostatními třídami, ale pravděpodobnost pro správné zařazení objektů mezi třídami **P** (prameny) a **HV** (hluboké vrty) bude zřejmě nižší.



Obr. 3 Grafy komponentních vah pro komponenty 1 a 2, 1 a 3, 2 a 3, a Cattelův indexový graf vlastních čísel ukazuje na počet využitelných hlavních komponent

(b) Korelace diskriminátorů: Dalším důležitým předpokladem pro diskriminační analýzu je neexistence multikolinearity v datech, což znamená, že dva a více diskriminátorů by neměly být silně korelovány. Jinak by totiž bylo možné predikovat jeden diskriminátor z jiného, což není vhodné zejména při užití krokové metody diskriminační analýzy. Z průzkumové analýzy EDA grafu komponentních vah na obr. 3 je zřejmé, že silná korelace byla indikována uvnitř trojice diskriminátorů *M-K-KNK4.5* a *M-K-SO4* a také uvnitř dvojice *K-Cl* a *A-CHSKMn*. Středně silná korelace je

uvnitř dvojic diskriminátorů *R-O2*, *A-PO4, M-Cl, CHSKMn-ZNK8.3, CHSKMn-PO4*. Korelační matice potvrzuje toto předběžné hodnocení dat a volbu diskriminátorů v předchozích dvou tabulkách. Sledované znaky *K, M, KNK4.5, SO4* a *Cl* lze nahradit jediným z důvodu silné korelace mezi nimi. Stejně tak dvojici znaků *A* a *CHSKMn* lze nahradit ze stejného důvodu jediným. U žádného z bodových grafů nelze pozorovat dělení mraku bodů na více shluků, tedy nelze předpokládat, že by mezi nimi byl diskriminátor s dobrou separační schopností objektů do tříd. Zároveň tvary většiny histogramů ukazují spíše na rozdělení log-normální. Dle korelační matice v tabulce 2 s nižšími korelacemi se jeví jako nejvhodnější diskriminátor *O2*, v histogramu je patrné bimodální rozdělení a v bodových grafech kombinací znaků *O2* s *pH* a *O2* s *ZNK8.3* se mrak bodů trhá na více shluků. Naproti tomu rozdělení znaků *NH4*, *NO2* a *NO3* vykazuje zřetelné zešikmení k nižším hodnotám, což vede k závěru o jejich nevhodnosti pro užití v metodě diskriminační analýzy. Vzhledem k silné korelaci mezi některými znaky a výraznému zešikmení rozdělení u některých dalších, je nutné zredukovat počet diskriminátorů. V dalším postupu diskriminační analýzy budeme klást důraz na pouze vybrané a účinné diskriminátory *K, CHSKMn, pH, T, R, O2, ZNK8.3* a *NO3*.

(c) Výstavba diskriminačního modelu: Vyšetření vlivu jednotlivých diskriminátorů přináší tabulka 3. V tabulce značí *Diskriminátor* jméno znaku. *Wilkovo λ při odstranění dotyčného diskriminátoru* udává hodnotu Wilkova kritéria λ vypočtenou při testování důsledku odstranění dotyčného diskriminátoru. Wilkovo kritérium λ vyjadřuje diskriminační sílu navrženého modelu. Jeho rozsah je od 1.0 se žádnou diskriminační silou až po 0.0 s perfektní diskriminační silou. *F-test při odstranění dotyčného diskriminátoru* představuje hodnotu *F*-kritéria vyčísleného k testování statistické významnosti Wilkova λ kritéria. *Spočtená hladina významnosti α při odstranění dotyčného diskriminátoru* je vypočtená hladina významnosti uvedeného *F*-testu při odstranění dotyčného diskriminátoru. Test je statisticky významný a diskriminátor je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$. Šest ze 16 vyšetřovaných diskriminátorů je v této úloze menší než 0.05, a proto jsou pro klasifikaci diskriminátorů do tříd statisticky významné a v úloze důležité. *Wilkovo λ pro dotyčný samotný diskriminátor* značí hodnotu Wilkova kritéria λ, kterou dostaneme za použití jediného diskriminátoru. *F-test pro dotyčný*

samotný diskriminátor představuje testační kritérium vyčíslené k testování statistické významnosti Wilkova λ kritéria. *Spočtená hladina významnosti α* se týká daného diskriminátoru. Uvedený *F*-test je statisticky významný a diskriminátor je pro klasifikaci znaků-diskriminátorů důležitý, je-li tato vypočtená hodnota α menší než uživatelem zadaná hladina významnosti α = 0.05. V tabulce 3 byla užita dopředná kroková analýza diskriminační analýzy a v šesti krocích byly nalezeny tyto znaky s dostatečnou diskriminační silou: *O2, K, R, CHSKMn, NO3* a *pH*.

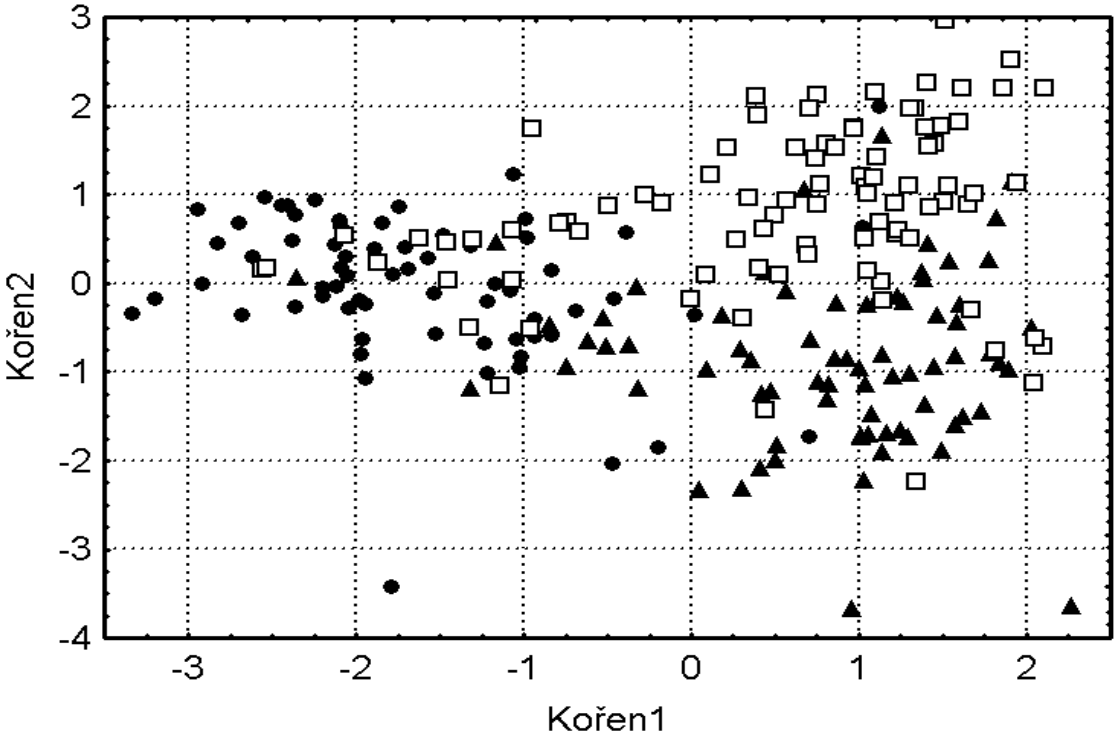
Tabulka 4 obsahuje právě stanovené odhady diskriminačních koeficientů klasifikační funkce, která slouží k zařazování nových objektů do tříd. Veličina *p* představuje apriorní pravděpodobnost, že bude objekt zařazen do dané třídy.

Ověření navržené diskriminační funkce se provádí klasifikací objektů, u nichž víme do jaké třídy patří v tak zvaném klasifikovaném výběru v tabulce 5. S ohledem na hodnoty procentuálního vyjádření správně zařazených objektů v jednotlivých třídách je zde navržený model nejméně účinný při klasifikaci objektů mělkých vrtů **MV**. Naopak nejméně chybně zařazených objektů je v třídě pramenů **P**.

(d) Diskriminační zařazování vzorků vod do tříd: Zařazování vzorků vody do tříd se odehrává na základě jejich Mahalanobisovy vzdálenosti. Vzdálenost se určuje mezi každým vzorkem vody a těžištěm každé třídy, definovaným jako průměr objektů ve třídě. Čím blíže je vzorek (objekt) umístěn k těžišti třídy, tím silnější je předpoklad, že vzorek patří do této třídy. Je možné také přímo vyčíslit pravděpodobnost, že vzorek patří do dané třídy. Jde o *posteriorní pravděpodobnost*. Aktuální klasifikace zobrazuje několik sloupců zařazení objektů, vzorků vody. Sloupce představují první, druhou a třetí možnost zařazení. Ve sloupci 1 tabulky 6 je nejvyšší posteriorní pravděpodobnost zařazení do správné třídy vzorků. Řádky označené hvězdičkou jsou chybně zařazené vorky. Znovu vidíme, že v této úloze je klasifikační správnost vysoká. Tabulka obsahuje Mahalanobisovy vzdálenosti klasifikovaných vzorků od *Z*-skóre jednotlivých tříd v tabulce 6. Vlivem toho, že shluky vzorků v rámci jednotlivých tříd se částečně prolínají, může docházet i k chybným zařazením vzorků vod z jedné třídy, které jsou blíže k centru (*Z*-skóre) třídy jiné.

Jiným způsobem klasifikace vzorků vod do tříd je využití hodnot aposteriorních pravděpodobností v tabulce 7. Vzorek vody je přidělen k té třídě, pro níž je hodnota

pravděpodobnosti co nejvyšší. Je zajímavé, že oba uvedené způsoby klasifikace vedly k chybnému zařazení týchž stejných vzorků, přestože u některých vyjímečných došlo k zařazení do odlišné nesprávné třídy. V řádku se u každého chybně zařazeného vzorku vody nachází vždy název známé a do výpočtu zadávané třídy vzorků vody a dále nalezené predikované třídy vzorků. Následuje hodnota pravděpodobnosti (v procentech), že se vzorek vody nachází v dané třídě vzorků. Hodnota blízko 100 % ukazuje, že vzorek skutečně patří do dotyčné třídy. Při užití lineární diskriminační techniky se vyčíslí pravděpodobnosti $P(i)$, že tento vzorek vody v řádku patří do i-té třídy. Nechť f_i , $i = 1, ..., k$, je hodnota lineární diskriminační funkce a $\max(f_k)$ je maximální diskriminační skóre ze všech tříd. Když užijeme regresní klasifikační techniku, bude $P(i)$ představovat predikovanou hodnotu regresní rovnice. Implicitně je y v regresní rovnici rovno 1 nebo 0 v závislosti na tom, zda objekt do i-té třídy vzorků patří či ne. Proto predikovaná hodnota blízko nuly ukazuje, že vzorek vody nepatří do i-té třídy, zatímco hodnota blízko 1 ukazuje na vysokou pravděpodobnost, že vzorek patří do i-té třídy. V žádném případě nemůže vyčíslená hodnota být větší než 1 a menší než 0.



Obr. 4 Graf lineárního diskriminačního skóre ukazuje na klasifikační zařazení jednotlivých vzorků podzemních vod do tří značně se překrývajících tříd

(e) Zařazení neznámých vzorků vody: Na základě diagramů skóre se snáze interpretují výsledky zařazení i neznámých vzorků vod v tabulce 8. Diagramy poskytují vizuální ověření, jak diskriminační funkce zařazují objekty do tříd. Předložený diagram ukazuje hodnoty prvního a druhého kanonického skóre. Je patrné, že již první kanonická funkce postačuje k zařazování vzorků vody, protože třídy vzorků vody mohou být snadno odděleny vertikální osou. Je možné také 3D zobrazení s průběžnou spojitou rotací tříd objektů podél jednotlivých os v prostoru, například v jazyce programu S-Plus. V takovém prostorovém zobrazení by bylo vytvoření a rozlišení tříd vzorků vody ještě názornější.

(f) Závěr: Všechna data z chemických analýz nebyla shledána jako vhodná pro dostatečně přesné přiřazení vzorků podzemních vod lineární diskriminační analýzou (LDA) do tří základních skupin. Zvláště procento správně zařazených vzorků vod v rámci třídy mělkých vrtů je dost nízké, pouze 58 %. Příčina může být jednak v tom, že monitorované ukazatele nemají dostatečnou diskriminační „sílu“ a také v tom, že většina diskriminátorů vykazuje jiné než normální rozdělení.

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

Doporučená literatura

[1] ČHMÚ – databáze jakosti vody (http://hydro.chmi.cz/ojv2/)

[2] STATISTICA 7.0 (http://www.statsoft.cz)

[3] Statistická ročenka životního prostředí České republiky 2006. MŽP ČR, Praha 2006.

[4] NCSS 2000 (http://www.ncss.com/)

[5] Meloun M., Militký J., Hill M.: Počítačová analýza vícerozměrných dat v příkladech. Academia, Praha 2005.

[6] Pytela O.: Chemometrie pro organické chemiky. Univerzita Pardubice, skripta, Pardubice 2003.

[7] Meloun M., Militký J.: Kompendium statistického zpracování experimentálních dat. Academia, Praha 2002 (1. vydání), 2006 (2. rozšířené vydání).

[8] Meloun M., Militký J.: Statistická analýza experimentálních dat. Academia, Praha 2004.