

Statistické zpracování vodohospodářských dat

10. Klasifikace podzemních vod diskriminační analýzou

Milan Meloun, Jindřich Freisleben

Klíčová slova

DA – PCA – CM – Cattelův graf – diskriminace – klasifikace – diskriminační skóre – Fisherova diskriminační funkce

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 podobnosti nebo rozdílu se pak provede klasifikace vzorků podzemních vod bud' čistě subjektivně na základě zkušeností, nebo objektivními metodami. Na základě diagramu 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.

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 výstupu v řeči v objekty zařazené do tříd. Účelem je nalézt prediktivní 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 nestejně 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 maticy 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 využít z rovnice pro aposteriorní pravděpodobnost příslušnosti k j -té skupině ($j = 1, 2$) a hledat její maximum. Podle typu hustot pravděpodobnosti pro znaky $f_j(x)$ a $f_i(x)$ se tak liší jednotlivé diskriminační metody v tom, jak jsou specifikovány dělicí oblasti a jaký májí 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 směsi 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

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

kde $m = 2$ je počet znaků.

2.3 Úprava prahového bodu

Dosud byl prahový bod C užíván jako bod, jenž udává stejně 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

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

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, \dots, 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 matic. Standardizované koeficienty se vypočítají 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 vícenásobné regresní analýzou. V diskriminační analýze se místo testování, zda se hodnota čtvrtce vícenásobného korelačního koeficientu R^2 změní přidáním nebo odebráním proměnného testuje, zda se změní hodnota Mahalanobisovy vzdálenosti D_M^2 . Obvykle se užívají popsána 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 krokovém 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, znova přečten, 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. Vybíráni diskriminátorů se ukončí, když žádné další diskriminátory nesplňují zaváděcí nebo odstraňovací kritérium.

2.5 Kritéria pro vybírání diskriminátorů

Existuje několik rozchodovacích 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 povolen jeden krok. Maximální počet kroků k vybírání diskriminátorů je roven dvojnásobku jejich počtu. Podobně jako ve vícenásobné regrese, 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 diskriminátem 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 a 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 obvyčejně vypočtena z rozdělení F , protože je zde vyšetřeno mnoho diskriminátorů a jsou vybrány největší a nejmenší hodnota F . Skutečnou hladinu významnosti α je obtížné vyčíslet, 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 rovně 1, protože v modelu dosud nejsou diskriminátory. Vedle Wilkova kritéria λ se pro statistickou významnost každého diskriminátoru vypočísluje také F -test. Hodnota F pro změnu Wilkova kritéria λ při přidání diskriminátoru do modelu tak, že model obsahuje celkem p diskriminátorů, se vypočíslí dle vztahu

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

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 kritéria.

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}_{ij} je průměr i -tého

diskriminátoru ve třídě 1. Protože je Mahalanobisova vzdálenost $D_{1,2}^2$ kritériem pro volbu diskriminátorů, je toto kritérium všech páru tříd vyčísleno jako první. Ten diskriminátor, který měl největší hodnotu $D_{1,2}^2$ pro dvě od začátku nejtěsnější třídy, čili které měly nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu.

3 Úloha: 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, představujících klasifikovaný soubor k vlastnímu otestování navrženého diskriminačního modelu. Za *závisle proměnnou* budeme brát znak TO , který značí typ objektu, nabývající tři hodnoty: 0 značí pramen, 1 značí mělký vrt, a konečně 2 značí hluboký vrt. Za *nezávisle proměnné* budeme považovat 16 diskriminátorů: pH značí naměřené pH vody v laboratoři, K je naměřená hodnota konduktivity [mS/m], T je teplota vody [$^{\circ}$ C], R je oxidačně redukční potenciál [mV], A absorbance změřená při vlnové délce 254nm v kyvetě délky 1 cm, M je celková mineralizace [mg/l], O_2 obsah kyslíku rozpustěného [mg/l], $CHSKMn$ chemická spotřeba kyslíku manganistanem [mg/l], $KNK4,5$ kyselinová neutralizační kapacita do pH 4,5 [mmol/l], $ZNK8,3$ 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].

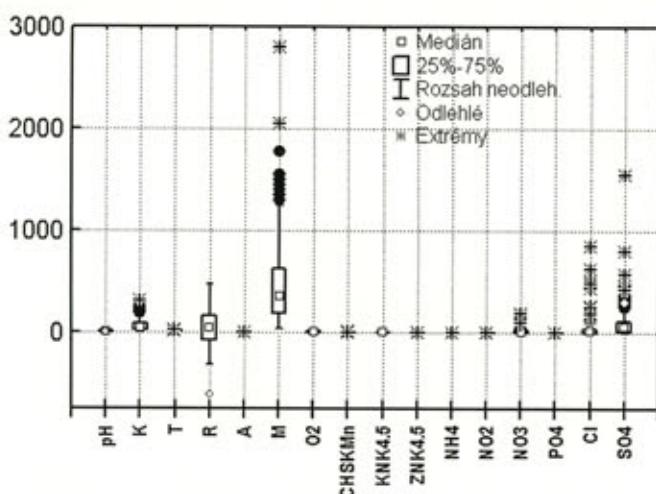
• Řešení:

(a) **Exploratorní analýza diskriminátorů:** Zdrojová matice dat analyzovaného výběru obsahuje 231 řádků vzorků vod a 16 sloupců vyšetřovaných diskriminátorů a neobsahuje žádné chybějící prvky. Zádný objekt v řádku není třeba vyřadit pro nedostatečný popis diskriminátorů. Byl použit statistický program STATISTICA 7.0 (StatSoft Praha). Nejprve se provede exploratorní analýza dat EDA a výčíslení popisných statistik (tabulka 1). 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 vyfazeny 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 šířnost rozdělení je blízká 0 a špičatost blízká 3. Z těchto hledisek není normalita splněna u většiny znaků. Nejmarkantnější odchyly od normality se objevují u diskriminátorů A , NH_4 , NO_2 , NO_3 , PO_4 , Cl a SO_4 . 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 SO_4 . 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 , $CHSKMn$, PO_4 . Zdá se, že lze docílit dobré separace mezi třídou MV (mělké vody) a ostatními třídami, ale pravděpodobnost pro správné zařazení objektů mezi třídami P (prameny) a HV (hluboké vody) bude zřejmě nižší.

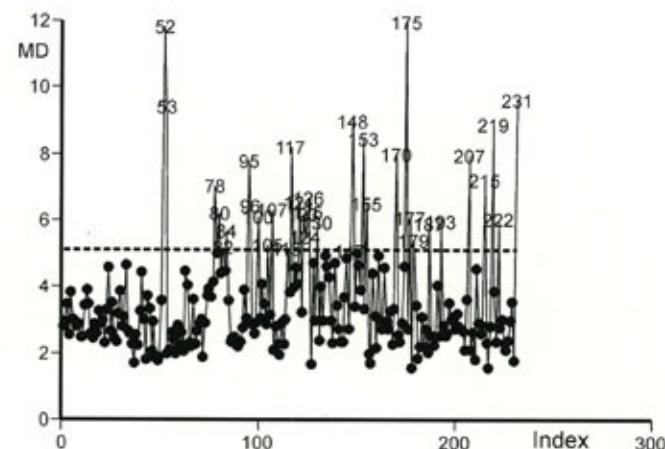
(b) **Korelace diskriminátorů:** Dalším důležitým předpokladem pro diskriminační analýzu je neexistence multikolinearity v datech, což znamená, že dva a více diskriminátorů by neměly být silně korelovány. Jinak by totiž bylo možné predikovat jeden diskrimi-

Tabulka 1. Základní popisné charakteristiky diskriminátorů ve třech třídách P (kód 0), MV (kód 1), HV (kód 2)

	TO	pH	K	T	R	A	M	O2	CHSKMn	KNK4.5	ZNK8.3	NH4	NO2	NO3	PO4	CI	SO4	
Aritmetický průměr	P	0	6,71	37,5	9,5	188	0,016	282	8,8	0,7	2,05	0,62	0,03	0,008	21,5	0,10	13,1	53,9
	MV	1	6,83	97,6	10,9	43	0,079	730	2,6	2,3	4,71	1,32	0,51	0,037	29,5	0,19	58,9	199,2
	HV	2	7,07	46,4	10,8	-25	0,028	344	3,5	0,9	3,30	0,77	0,26	0,015	11,1	0,11	23,8	38,3
Směrodatná odchylka	P	0	0,73	29,7	1,2	115	0,015	232	2,0	0,6	2,18	0,51	0,04	0,028	29,3	0,11	13,6	48,1
	MV	1	0,50	66,0	1,7	123	0,074	579	2,2	2,0	3,87	0,84	1,11	0,055	40,0	0,37	48,8	288,8
	HV	2	0,73	38,9	2,8	119	0,047	232	3,3	1,0	2,24	0,65	0,74	0,025	26,3	0,16	57,4	70,0
Medián	P	0	6,97	28,0	9,5	185	0,013	205	9,2	0,5	0,81	0,51	0,03	0,003	11,5	0,06	6,5	42,3
	MV	1	6,97	77,8	10,6	46	0,055	580	2,0	1,7	4,61	1,14	0,09	0,014	5,8	0,08	49,0	109,5
	HV	2	7,17	35,8	10,1	-28	0,015	285	2,0	0,5	3,03	0,55	0,05	0,005	0,5	0,04	8,4	20,5
Šíkmost	P	0	-0,3	1,0	0,1	3,2	2,6	1,2	-1,4	2,7	1,0	2,1	5,2	6,0	3,2	2,4	1,5	2,3
	MV	1	-1,4	2,0	2,6	0,4	2,6	2,5	1,5	1,8	3,7	1,1	4,4	3,3	1,4	4,5	1,7	4,2
	HV	2	-0,2	2,1	2,9	0,0	6,0	1,8	0,9	2,6	0,7	0,9	4,8	3,4	3,8	2,9	4,8	5,4
Špičatost	P	0	-1,3	0,3	-0,5	20,4	7,3	1,3	2,9	8,4	-0,3	4,9	31,5	36,3	14,2	7,3	1,5	6,9
	MV	1	2,5	5,1	8,8	-0,3	9,9	8,4	2,2	3,8	22,3	0,8	24,5	14,7	0,7	22,7	4,6	20,6
	HV	2	0,5	5,1	10,4	0,0	45,0	5,9	-0,6	6,8	0,3	-0,3	23,9	14,1	18,1	8,8	23,6	34,8
Počet objektů	P	0	69	69	69	69	69	69	69	69	69	69	69	69	69	69	69	
	MV	1	73	73	73	73	73	73	73	73	73	73	73	73	73	73	73	
	HV	2	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	

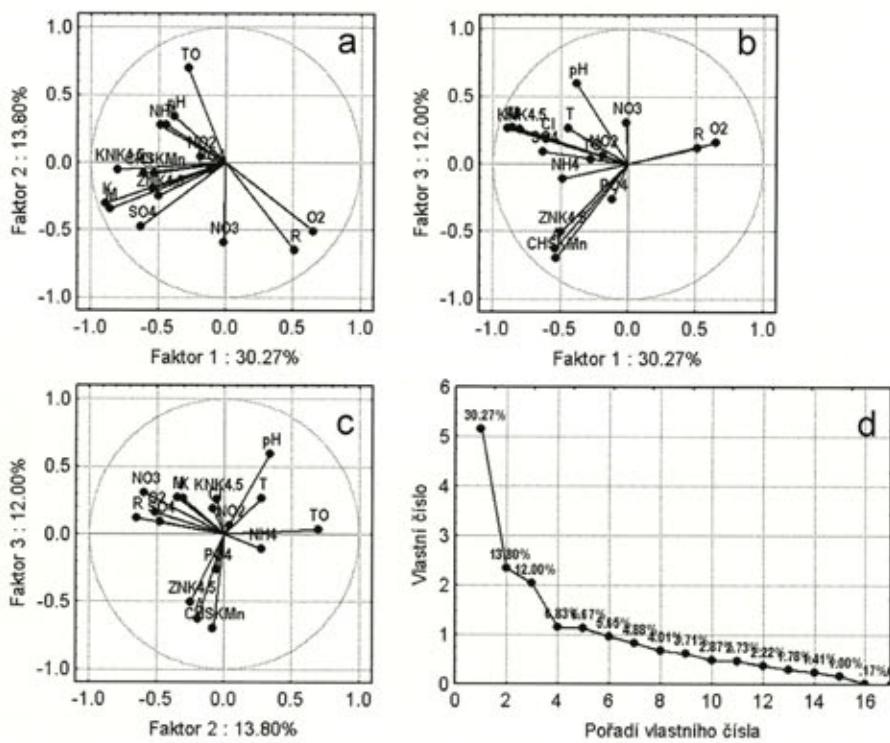


Obr. 1. Krabicové grafy všech diskriminátorů ukazují na míru rozptýlení a proměnlivost



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

nátor 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 $M-KN4.5$ a $M-K-SO4$ a také uvnitř dvojice $K-CI$ a $A-CHSKMn$. Středně silná korelace je uvnitř dvojic diskriminátorů $R-O2$, $A-PO4$, $M-CI$, $CHSKMn-ZNK8.3$, $CHSKMn-PO4$. Korelační matici potvrzuje toto předběžné hodnocení dat a volbu diskriminátorů v předchozích dvou tabulkách. Sledované znaky K , M , $KN4.5$, $SO4$ a CI 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 matici v tabulce 2 s nižšími korelacemi se jeví jako nevhodnější 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



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

Tabulka 2. Korelační matici diskriminátorů. Tučné je vyznačen statisticky významný Pearsonův korelační koeficient.

	pH	K	T	R	A	M	O2	CHSKMn	KNK4.5	ZNK8.3	NH4	NO2	NO3	PO4	Cl	SO4
pH	1,00															
K	0,34	1,00														
T	0,33	0,32	1,00													
R	-0,32	-0,23	-0,28	1,00												
A	-0,08	0,35	0,04	-0,19	1,00											
M	0,32	0,95	0,30	-0,19	0,33	1,00										
O2	-0,25	-0,36	-0,37	0,62	-0,31	-0,37	1,00									
CHSKMn	-0,13	0,31	0,05	-0,28	0,75	0,30	-0,36	1,00								
KNK4.5	0,52	0,76	0,35	-0,33	0,32	0,76	-0,47	0,20	1,00							
ZNK8.3	-0,29	0,36	0,10	-0,17	0,47	0,33	-0,32	0,50	0,36	1,00						
NH4	0,23	0,34	0,19	-0,36	0,31	0,25	-0,30	0,33	0,23	0,07	1,00					
NO2	0,12	0,14	0,02	-0,06	0,06	0,11	-0,18	0,03	0,14	0,06	0,04	1,00				
NO3	-0,02	0,23	-0,00	0,35	-0,02	0,23	0,26	-0,13	0,08	0,06	-0,19	0,07	1,00			
PO4	0,03	0,04	0,04	-0,04	0,26	0,01	-0,11	0,19	0,08	0,07	0,01	0,11	0,05	1,00		
Cl	0,19	0,75	0,20	-0,22	0,18	0,60	-0,25	0,22	0,32	0,15	0,37	0,08	0,06	0,02	1,00	
SO4	0,08	0,70	0,18	-0,05	0,30	0,80	-0,17	0,28	0,49	0,38	0,11	0,09	0,16	-0,03	0,20	1,00

o jejich nevhodnosti pro užití v metodě diskriminační analýzy. Vzhledem k silné korelace 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 tabulku 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 a 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 zadáná 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 a se týká daného diskriminátoru. Uvedený F-test je statisticky významný a diskriminátor je pro klasifikaci znaků-diskriminátorů důležitý, je-li tato vypočtená hodnota a menší než uživatelem zadáná hladina významnosti $\alpha = 0,05$. 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 chyběně zařazených objektů je v třídě pramenů P.

(d) **Diskriminační zařazování vzorků podle 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

Tabulka 3. Testační kritéria po krokovém vyhledání vhodných diskriminátorů. Počet testovaných diskriminátorů v modelu je 16. Grupovací proměnnou je: TO (na 3 třídy). Wilkovo kritérium lambda: 0,31258 přibliž F (32,426) = 10,499 p < 0,0000. Tučné jsou vyznačeny statisticky významné diskriminátory

	Wilkovo lambda	Parc. lambda	F na vyj [2,223]	Úroveň p	Toler.	1-toler. R ^ 2
pH	0,312811	0,999258	0,07908	0,923991	0,417522	0,582479
K	0,323925	0,964974	3,86568	0,022434	0,005894	0,994106
T	0,316463	0,987726	1,32342	0,268403	0,844840	0,155160
R	0,333681	0,936761	7,18964	0,000952	0,651227	0,348773
A	0,323661	0,965761	3,77574	0,024469	0,432417	0,567583
M	0,313324	0,997622	0,25387	0,776031	0,050159	0,949841
O2	0,402291	0,776998	30,56604	0,000000	0,598848	0,401152
CHSKMn	0,315030	0,992219	0,83515	0,435223	0,371028	0,628972
KNK4.5	0,320485	0,975329	2,69391	0,069920	0,035599	0,964401
ZNK4.5	0,322249	0,969991	3,29479	0,038974	0,441284	0,558716
NH4	0,313962	0,995595	0,47120	0,624899	0,638794	0,361206
NO2	0,315708	0,990089	1,06609	0,346182	0,924779	0,075221
NO3	0,316548	0,987462	1,35228	0,260861	0,277390	0,722611
PO4	0,320713	0,974637	2,77143	0,064830	0,889275	0,110725
Cl	0,324119	0,964396	3,93184	0,021047	0,017028	0,982972
SO4	0,321440	0,972434	3,01897	0,050947	0,029907	0,970093

Tabulka 4. Klasifikační funkce: apriorní pravděpodobnost p, že objekt bude zařazen do dotyčné třídy

G_1:0 (prameny) p = 0,29870	G_2:1 (mělké vrt)	G_3:2 (hluboké vrt) p = 0,32035
pH	35,608	35,592
K	0,273	0,379
T	2,516	2,686
R	0,030	0,027
A	72,953	83,856
M	0,040	0,041
O2	2,439	1,675
CHSKMn	0,226	-0,136
KNK4.5	-10,917	-11,861
ZNK4.5	19,578	20,873
NH4	-2,183	-2,328
NO2	0,793	6,208
NO3	-0,162	-0,161
PO4	4,400	3,929
Cl	-0,151	-0,185
SO4	-0,095	-0,111
konst.	-146,509	-146,300

Tabulka 5. Klasifikační matici: procentuelní vyjádření správně zařazených objektů

% správně zařazených objektů	G_1:0 (prameny)	G_2:1 (mělké vrt)	G_3:2 (hluboké vrt)
G_1:0 (prameny)	79,71	55	6
G_2:1 (mělké vrt)	58,11	6	43
G_3:2 (hluboké vrt)	73,86	13	10
Celkem	70,56	74	59

Tabulka 6. Přehled chybně zařazených objektů dle kritéria Mahalanobisovy vzdálenosti. Tučně je vyznačeno chybné zařazení proti původně zadanému

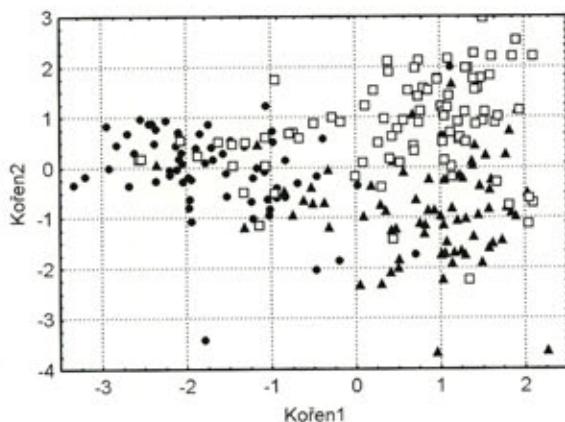
Typ objektu	Prameny	Mělké vrty	Hluboké vrty	Typ objektu	Prameny	Mělké vrty	Hluboké vrty		
PO0016	pramen	9,698	4,786	1,644	VP0635	mělký vrt	6,399	3,417	2,490
PO1835	pramen	17,363	9,850	4,089	VP0651	mělký vrt	9,942	4,876	4,802
PO4006	pramen	8,249	14,974	7,015	VP0672	mělký vrt	8,421	8,126	6,412
PP0053	pramen	9,737	11,258	5,708	VP0692	mělký vrt	6,148	3,436	3,533
PP0111	pramen	14,104	5,417	10,392	VP0714	mělký vrt	3,736	16,204	11,627
PP0115	pramen	5,296	9,733	2,998	VP1324	mělký vrt	6,173	9,011	10,633
PP0148	pramen	5,252	6,905	4,942	VP1576	mělký vrt	10,079	3,321	3,441
PP0331	pramen	17,615	12,406	16,491	VP1601	mělký vrt	4,721	4,943	6,788
PP0378	pramen	14,869	11,176	13,815	VP1721	mělký vrt	5,245	5,807	7,768
PP0434	pramen	5,856	2,948	6,044	VP1724	mělký vrt	9,828	3,860	3,690
PP0462	pramen	3,692	4,851	1,088	VP1942	mělký vrt	3,181	8,175	5,416
PP0496	pramen	38,041	33,869	45,964	VP1966	mělký vrt	7,935	5,487	0,952
PP0513	pramen	33,262	16,923	31,375	VP7005	hluboký vrt	13,630	7,218	8,409
PP0540	pramen	5,642	5,791	4,272	VP7012	hluboký vrt	4,010	8,956	5,321
VB0031	mělký vrt	9,320	2,022	1,715	VP7016	hluboký vrt	3,042	10,424	5,187
VB0078	mělký vrt	6,185	2,706	2,861	VP7215	hluboký vrt	30,313	16,356	20,667
VB0095	mělký vrt	10,635	2,474	2,592	VP7304	hluboký vrt	3,645	8,186	4,267
VB0117	mělký vrt	8,789	3,808	2,155	VP7510	hluboký vrt	5,368	13,504	8,337
VB0236	mělký vrt	10,151	4,194	3,842	VP7524	hluboký vrt	2,309	5,912	4,914
VB0271	mělký vrt	11,236	2,662	2,874	VP7614	hluboký vrt	21,540	8,523	12,936
VB0295	mělký vrt	9,757	2,121	2,393	VP7618	hluboký vrt	36,057	20,343	28,961
VB0322	mělký vrt	11,183	3,914	1,626	VP7710	hluboký vrt	12,251	6,125	11,142
VB0401	mělký vrt	1,646	5,213	4,547	VP7716	hluboký vrt	4,634	11,720	9,472
VB9754	hluboký vrt	33,323	15,267	24,906	VP7718	hluboký vrt	7,515	9,895	11,220
VO0005	mělký vrt	12,131	4,289	2,474	VP7720	hluboký vrt	3,920	13,556	9,810
VO0016	mělký vrt	9,205	3,652	3,991	VP7722	hluboký vrt	1,774	15,629	11,226
VO0089	mělký vrt	14,551	10,366	4,293	VP7727	hluboký vrt	2,484	7,468	4,257
VP0007	mělký vrt	13,070	4,038	2,020	VP7800	hluboký vrt	14,021	7,638	9,839
VP0094	mělký vrt	4,121	2,745	2,811	VP8206	hluboký vrt	22,375	9,310	20,565
VP0119	mělký vrt	4,004	5,373	4,276	VP8419	hluboký vrt	29,351	16,416	20,705
VP0210	mělký vrt	12,544	3,138	2,518	VP8456	hluboký vrt	2,099	12,261	7,926
VP0326	mělký vrt	11,529	1,909	2,186	VP8503	hluboký vrt	6,447	17,620	17,062
VP0478	mělký vrt	12,811	2,879	2,488	VP9500	hluboký vrt	2,145	8,778	3,702
VP0485	mělký vrt	48,752	37,122	31,262	VP9506	hluboký vrt	45,671	28,198	38,944

Tabulka 7. Přehled chybně zařazených objektů dle kritéria aposteriorní pravděpodobnosti. Tučně je vyznačeno chybné zařazení proti původně zadanému

Typ objektu	Prameny	Mělké vrty	Hluboké vrty	Typ objektu	Prameny	Mělké vrty	Hluboké vrty		
PO0016	pramen	0,0117	0,1439	0,8444	VP0635	mělký vrt	0,0675	0,3175	0,6150
PO1835	pramen	0,0010	0,0439	0,9551	VP0651	mělký vrt	0,0321	0,4273	0,5406
PO4006	pramen	0,2918	0,0107	0,6975	VP0672	mělký vrt	0,1740	0,2133	0,6128
PP0053	pramen	0,0896	0,0443	0,8661	VP0692	mělký vrt	0,1013	0,4159	0,4828
PP0111	pramen	0,0110	0,8980	0,0910	VP0714	mělký vrt	0,9737	0,0020	0,0243
PP0115	pramen	0,1929	0,0222	0,7849	VP1324	mělký vrt	0,7170	0,1835	0,0994
PP0148	pramen	0,3368	0,1559	0,5073	VP1576	mělký vrt	0,0148	0,4587	0,5265
PP0331	pramen	0,0569	0,8143	0,1288	VP1601	mělký vrt	0,4157	0,3935	0,1907
PP0378	pramen	0,1011	0,6780	0,2209	VP1721	mělký vrt	0,4620	0,3691	0,1688
PP0434	pramen	0,1492	0,6757	0,1751	VP1724	mělký vrt	0,0201	0,4210	0,5589
PP0462	pramen	0,1579	0,0936	0,7485	VP1942	mělký vrt	0,6627	0,0577	0,2796
PP0496	pramen	0,1048	0,8926	0,0026	VP1966	mělký vrt	0,0213	0,0766	0,9020
PP0513	pramen	0,0003	0,9988	0,0009	VP7005	hluboký vrt	0,0224	0,5847	0,3929
PP0540	pramen	0,2203	0,2162	0,5635	VP7012	hluboký vrt	0,5686	0,0507	0,3807
VB0031	mělký vrt	0,0101	0,4087	0,5812	VP7016	hluboký vrt	0,6814	0,0180	0,3006
VB0078	mělký vrt	0,0723	0,4358	0,4919	VP7215	hluboký vrt	0,0008	0,8755	0,1237
VB0095	mělký vrt	0,0074	0,4618	0,5308	VP7304	hluboký vrt	0,4867	0,0532	0,4601
VB0117	mělký vrt	0,0203	0,2587	0,7210	VP7510	hluboký vrt	0,7631	0,0138	0,2231
VB0236	mělký vrt	0,0192	0,3997	0,5811	VP7524	hluboký vrt	0,6556	0,1145	0,2299
VB0271	mělký vrt	0,0062	0,4739	0,5199	VP7614	hluboký vrt	0,0012	0,8806	0,1182
VB0295	mělký vrt	0,0100	0,4796	0,5104	VP7618	hluboký vrt	0,0004	0,9835	0,0161
VB0322	mělký vrt	0,0051	0,2061	0,7888	VP7710	hluboký vrt	0,0386	0,8746	0,0868
VB0401	mělký vrt	0,6756	0,1201	0,2043	VP7716	hluboký vrt	0,8731	0,0267	0,1002
VB9754	hluboký vrt	0,0001	0,9901	0,0097	VP7718	hluboký vrt	0,6561	0,2112	0,1327
VO0005	mělký vrt	0,0046	0,2475	0,7478	VP7720	hluboký vrt	0,9290	0,0079	0,0630
VO0016	mělký vrt	0,0282	0,4790	0,4928	VP7722	hluboký vrt	0,9877	0,0010	0,0113
VO0089	mělký vrt	0,0044	0,0377	0,9579	VP7727	hluboký vrt	0,6177	0,0541	0,3283
VP0007	mělký vrt	0,0024	0,2296	0,7680	VP7800	hluboký vrt	0,0269	0,6923	0,2808
VP0094	mělký vrt	0,1790	0,3767	0,4444	VP8206	hluboký vrt	0,0014	0,9943	0,0044
VP0119	mělký vrt	0,3760	0,2007	0,4234	VP8419	hluboký vrt	0,0013	0,8739	0,1248
VP0210	mělký vrt	0,0032	0,3744	0,6224	VP8456	hluboký vrt	0,9289	0,0061	0,0650
VP0326	mělký vrt	0,0037	0,4832	0,5131	VP8503	hluboký vrt	0,9898	0,0039	0,0063
VP0478	mělký vrt	0,0026	0,4018	0,5956	VP9500	hluboký vrt	0,6133	0,0235	0,3631
VP0485	mělký vrt	0,0001	0,0420	0,9579	VP9506	hluboký vrt	0,0002	0,9942	0,0056

Tabulka 8. Klasifikace prvních 14 neznámých objektů dle Mahalanobisovy vzdálenosti od těžiště třídy

Klasifikace	Prameny	Mělké vrtý	Hluboké vrtý
1	pramen	3,627	13,512
2	pramen	2,910	15,066
3	pramen	3,070	12,565
4	pramen	1,486	13,679
5	pramen	1,191	6,951
6	hluboký vrt	8,061	2,449
7	hluboký vrt	7,190	2,296
8	mělký vrt	23,034	10,948
9	mělký vrt	8,530	2,846
10	mělký vrt	8,356	2,820
11	hluboký vrt	10,700	3,338
12	mělký vrt	21,231	9,151
13	hluboký vrt	7,299	4,007
14	mělký vrt	15,186	8,192



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

předpoklad, že vzorek patří do této třídy. Je možné také přímo vyčís lit 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 tučně jsou chybně zařazené vory. Znovu vidíme, že v této úloze je klasifikační správnost vysoká. Tabulka obsahuje Mahalanobisovu vzdálenost klasifikovaných vzorků od Z-skóre jednotlivých tříd v tabulce 6. Vlivem toho, že shluhy 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 výjimečných došlo k zařazení do odlišné nesprávné třídy. V řádku se u každého chybně zařazeného vzorku voda nachází vždy název známé a do výpočtu zadávané třídy vzorků voda a dále nalezené predikované třídy vzorků. Následuje hodnota pravděpodobnosti (v procentech), že se vzorek voda 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 , $i = 1, \dots, k$, 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 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 vyčíslena hodnota být větší než 1 a menší než 0.

(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 hod-

noty 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 úlohy: Všechna data z chemických analýz nebyla shledána jako vhodná pro dostatečně přesné přiřazení vzorků podzemních vod lineární diskriminační analýzou (LDA) do tří základních skupin (obr. 4.). 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.

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

prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Fakulta chemicko-technologická
Univerzita Pardubice
Čs. Legii 565, 532 10 Pardubice
<http://meloun.upce.cz>
tel.: 466 037 026
e-mail: milan.meloun@upce.cz

Ing. Jindřich Freisleben,
Český hydrometeorologický ústav,
Na Šabatce 17, 143 06 Praha 4 – Komořany
tel.: 244 032 331
e-mail: freisleben@chmi.cz

Computer-Assisted Statistical Data Analysis. 10. Classification of underground water using a discriminant analysis (Meloun M., Freisleben J.)

Key words

DA – PCA – Cattel's graf – discriminant analysis – discriminant score – supervised learning – training set – Fisher discriminant function

The linear discriminant analysis enables classification among two or more groups of objects being described with more variables. The groups are known a priori and the aim is to devise rules which can allocate previously unclassified objects or individuals into these groups in an optimal fashion. The investigator has one set of multivariate observations, the training set, for which group membership is known with certainty a priori, and a second set, the test set, consisting of observations for which group membership is unknown and which have to be assigned to one of the known groups as accurately as possible. The information used in deriving a suitable allocation rule is the variable values of the training sample. Areas where this type of classification problem is of importance are numerous. To illustrate the application of Fisher's linear discriminant function the sample data of underground water was used to classify samples into three various source classes. Percentage of truly classified samples, however, is not high, 58% only, because the measured variables have low discriminant power for an efficient classification and mostly exhibit non-normal distribution. The software-assisted procedure of discriminant analysis is proposed and applied.