

Aplikace analýzy hlavních komponent pro redukci dimenze transportně-reakčního problému*

Jan Šembera

Ústav nových technologií a aplikované informatiky

Fakulta mechatroniky a mezioborových inženýrských studií Technické univerzity v Liberci

e-mail: jan.sembera@tul.cz

Abstrakt

Příspěvek se zabývá redukcí dimenze transportně-reakčního problému v této oblasti ne-standardním postupem. Je při něm využíván postup hojně aplikovaný v analýze signálu i při řešení dalších technických problémů – analýza hlavních komponent. Příspěvek nepřináší ucelenou metodiku redukce dimenze obecné úlohy s užitím analýzy hlavních komponent, ale na jednom příkladu ukazuje možnosti tohoto postupu a nastoluje řadu otázek, které je třeba řešit pro každou úlohu zvlášť.

1 Úvod

Výsledky popisované v této kapitole vycházejí z ročníkových prací autorových studentů Ivana Bruského a Bc. Lukáše Zedka a jsou částečně publikovány ve zprávách z jejich ročníkových prací [1] a [2].

Základním problémem transportních úloh zásadně ovlivňovaných chemickými reakcemi je příliš velká dimenze úlohy. Nechme nyní stranou, jak velká dimenze je příliš velká a jaká je již přijatelná (závislost těchto pojmů na konkrétní úloze a jejím uspořádání je skutečně významná). Základním postupem, který chemici při tvorbě chemického modelu užívají, je klasifikace složek roztoku na řídicí a marginální a vyčlenění hlavních chemických reakcí mezi řídicími složkami, které ovlivňují řešení problému zásadně, od ostatních chemických reakcí, které mohou být v modelu považovány za doplňující, nebo mohou být úplně vynechány. Tímto postupem je často výrazně redukována dimenze řešené úlohy s tím, že hlavní jevy a koncentrace řídicích složek v chemickém systému jsou simulovány a ostatní jevy a koncentrace lze v případě potřeby s větší či menší přesností následně odvozovat z výsledku simulace.

Tento postup je velmi výhodný v případě, že sledujeme vývoj koncentrací řídicích složek, nebo změny podmínek ovlivněných těmito řídicími složkami. Pokud nás zajímají některé jevy významně ovlivněné konkrétní marginální složkou, můžeme model rozšířit o tuto složku (zařadit ji mezi řídicí složky) a zvýšit dimenzi úlohy. Pokud ale studujeme takovou situaci, kdy pro nás všechny složky systému mají přibližně stejný význam, nelze tímto postupem účinně redukovat dimenzi úlohy jinak než na úkor kvality výsledku.

Taková situace nastala při modelování předpovědi dlouhodobého vývoje kontaminace na lokalitě Stráž pod Ralskem s. p. DIAMO po provedení sanace metodou neutralizace in-situ. V podzemí je směs roztoků s 22 měřenými složkami, z nichž některé přímo řídí hlavní chemické děje, proto je model nesmí opomíjet, ale mezi marginálními složkami jsou také nejnebezpečnější

*Tento výsledek byl realizován s podporou státního rozpočtu České republiky prostřednictvím projektu č. 1M0554 Výzkumné centrum Pokročilé sanační technologie a procesy v programu MŠMT Výzkumná centra (PP2-DP01) a s podporou Grantové agentury České republiky, projekt č. 102/06/P450.

kontaminanty. Ty model také nesmí opominout, protože jejich bilanci a šíření je třeba počítat co nejpřesněji.

Je tedy třeba zachovat počet simulovaných složek a zároveň redukovat dimenzi problému. To lze provést postupem vycházejícím z lineární algebry. Pohlízejme na množinu všech provedených analýz roztoků ve sledované lokalitě jako na množinu M vektorů ve 22-rozměrném lineárním vektorovém prostoru V , jehož souřadné osy odpovídají koncentraci jednotlivých složek roztoku. Hledejme takový n -rozměrný lineární vektorový podprostor V_n prostoru V , který bude „nejblíže“ množině M , tj. bude minimalizovat chybu průmětu E_n^2 definovanou jako součet druhých mocnin vzdáleností všech vektorů z M od jejich průmětů do V_n :

$$E_n^2 = \sum_{\mathbf{x} \in M} \|\mathbf{x} - \Pi_{V_n} \mathbf{x}\|^2. \quad (1)$$

Zde Π_{V_n} označuje operátor kolmého promítání do prostoru V_n .

Pokud se nám podaří najít podprostor dostatečně malé dimenze n_0 s dostatečně malou chybou průmětu $E_{n_0}^2$, můžeme zredukovat dimenzi úlohy transportu z původních 22 na n_0 a ze simulace chemických reakcí nevyčleňovat žádnou složku.

2 Úloha redukce dimenze

Problémem, který potřebujeme vyřešit, je nalezení vektorového prostoru V_n dané dimenze n , který je nejblíže množině dat M ve smyslu popsáném v předešlém odstavci. Tato úloha je ekvivalentní úloze najít ortogonální doplněk k prostoru V_n (označme ho D_n). Pokud dimenzi prostoru V označíme s (pro naši úlohu je $s = 22$), má D_n dimenzi $s - n$.

Pokud uvažujeme dál touto cestou, přijdeme k úvaze, že vůbec nejlepším východiskem pro řešení takové úlohy by bylo mít k dispozici ortogonální bázi prostoru V takovou, že první vektor báze bude „nejméně důležitý“ vzhledem k množině M , tj. průměty do nadroviny kolmé na tento vektor budou ze všech možných výběrů minimální. Druhý vektor této báze bude vybrán ze „zbytku prostoru V “, tj. též na něj kolmé nadroviny, a bude tedy kolmý na první vektor, a bude mít tutéž vlastnost vzhledem k uvedené nadrovině. A tak dále. Taková báze bude mít dále tu výhodu, že (protože bude ortogonální), kolmé promítání do jakéhokoli podprostoru generovaného některými jejími vektory bude spočívat pouze ve vynulování souřadnic v ostatních směrech.

Tento poněkud nepřehledný popis zformalizujeme následujícím způsobem:

- Nejprve hledejme první vektor ξ_s takto: $\xi_s = \arg \min_{\|y\|=1} \sum_{x \in M} x \cdot y$
- Potom hledejme druhý vektor ξ_{s-1} takto: $\xi_{s-1} = \arg \min_{\|y\|=1, y \perp \xi_s} \sum_{x \in M} x \cdot y$
- Potom hledejme další vektory ξ_{s-i} ($i = 2, \dots, n-1$) takto:

$$\xi_{s-i} = \arg \min_{\|y\|=1, y \perp \{\xi_s, \xi_{s-1}, \dots, \xi_{s-i+1}\}} \sum_{x \in M} x \cdot y$$

Takto získaná ortonormální báze bude mít tu vlastnost, že pro zvolené n bude optimální prostor V_n generován vektory této báze $\xi_1, \dots, \xi_n$.

Při hledání uvedených vektorů můžeme postupovat podle navrženého algoritmu, jehož jednotlivé kroky jsou samostatné optimalizační úlohy. Každá z nich má řešení, které není jednoznačné. Vždy existují alespoň dva vektory minimalizující účelovou funkci. Pokud jsou řešení právě dvě, je lhostejné, které řešení vybereme, protože oba vektory generují stejný podprostor (mají pouze opačný směr). Pokud má některá dílčí úloha více než dvě řešení, je už původní úloha nejednoznačná a pro účely praktického použití lze zvolit libovolné z přípustných řešení.

Než se pustíme do řešení optimalizačních úloh, zaměříme se na zdánlivě odtaziště téma – metodu analýzy hlavních komponent.

3 Analýza hlavních komponent

Analýza hlavních komponent (Principal Component Analysis - PCA) je metodou redukce dimenze s minimální ztrátou informace v datech standardně užívanou pro řešení řady technických problémů. Uplatňuje se také v ekonomických vědách a lékařství (např. [3, 4, 5]). Je založena na transformaci souřadného systému - nalezení speciální ortonormální báze prostoru, ve kterém jsou data umístěna. Vektory hledané ortonormální báze jsou uspořádány tak, že první určuje směr obsahující největší možnou jednorozměrnou informaci v datech a ve směru posledního bázevého vektoru je obsah informace v datech minimální. Tento postup se standardně užívá při zpracování signálu k de Korelaci dat.

Analýza hlavních komponent (viz např. [6]) je realizována následujícím postupem:

1. Uspořádá data do matice $\mathbf{X}$ typu $r \times s$. Každý řádek obsahuje jedno z r pozorování a sloupce odpovídají s měřeným veličinám.
2. Vypočítá průměrný vektor dat (průměrný řádek matice $\mathbf{X}$) $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_s)^T$, $\bar{x}_i = \frac{1}{r} \sum_{j=1}^r X_{ji}$ (zde X_{ji} značí prvek matice $\mathbf{X}$ v j -tém řádku a i -tém, sloupci) a vytvoří matici $\mathbf{X}^* = \mathbf{X} - \bar{\mathbf{1}} \cdot \bar{\mathbf{x}}^T$, kde $\bar{\mathbf{1}} = (1, \dots, 1)^T$ označuje sloupcový vektor délky r s jedničkovými prvky.
3. Vypočítá kovarianční matici $\mathbf{C} = \frac{1}{r-1} \mathbf{X}^{*T} \cdot \mathbf{X}^*$.
4. Spočítá vlastní čísla a vlastní vektory matice $\mathbf{C}$ (označme vlastní čísla uspořádaná v absolutní hodnotě od největšího k nejmenšímu λ_i a jim příslušné vektory ξ_i) a sestaví transformační matici $\mathbf{T}$ typu $r \times s$ obsahující vlastní vektor kovarianční matice $\mathbf{C}$ ve sloupcích (tj. $T_{ij} = (\xi_j)_i$).
5. Vybere n hlavních komponent $\xi_1, \dots, \xi_n$ a sestaví transformační matici $\mathbf{T}_n$ typu $r \times n$ obsahující prvních n vlastních vektorů kovarianční matice $\mathbf{C}$ ve sloupcích (tj. $T_{n,ij} = (\xi_j)_i$).
6. Redukuje původní data (ortogonálně je promítne do podprostoru generovaného hlavními komponentami) $\mathbf{Z}_n = \mathbf{X}^* \mathbf{T}_n$.
7. Rekonstruuje redukovaná vycentrovaná data $\mathbf{Y}^* = \mathbf{Z}_n \mathbf{T}_n^T$.
8. Rekonstruuje redukovaná původní data $\mathbf{Y} = \mathbf{Y}^* + \bar{\mathbf{1}} \cdot \bar{\mathbf{x}}^T$.

Matice $\mathbf{Y}$ pak obsahuje původní data kolmo promítnutá do afinního podprostoru dimenze n , který je nejlepší v tom smyslu, že celková chyba E_n^2 (1) způsobená promítáním je minimální.

Jednoduchost promítání v kroku 6 je založena na tom, že kovarianční matice $\mathbf{C}$ je pro každou sadu dat symetrická pozitivně semidefinitní a tedy její vlastní vektory tvoří ortonormální systém. Matice $\mathbf{T}$ je tedy ortogonální a její inverze je totožná s její transpozicí, podobně pseudoinverzní matice k $\mathbf{T}_n$ je totožná s její transpozicí.

Kroky 2 a 8 jsou prováděny pouze pro účely získání přesnějších aproximací u nevycentrovaných dat a princip metody s nimi nijak nesouvisí. Stejně tedy bude algoritmus fungovat, pokud je nahradíme kroky

2a. $\mathbf{X}^* = \mathbf{X}$.

8a. $\mathbf{Y} = \mathbf{Y}^*$.

Nebude tak nalezen nejlepší afinní podprostor, ale nejlepší lineární podprostor ve stejném smyslu.

Zde je třeba si povšimnout, že po navržené úpravě jde o hledání právě těch směrů, které byly popsány ve formulaci úlohy redukce dimenze popsané v odstavci 2, jen jsou všechny směry nalezeny najednou. V následujícím textu se tedy nebudeme zabývat realizací postupu z odstavce 2, ale zaměříme se na výsledky získané analýzou hlavních komponent.

n	21	20	19	18	17	16	15	14
E_n^2	0,135	25,8	60,0	155	291	559	1432	3110
E_n	0,367	5,08	7,75	12,5	17,1	23,6	37,8	55,8
p_{E_n}	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%

n	13	12	11	10	9	8
E_n^2	5983	$1,12 \cdot 10^4$	$2,17 \cdot 10^4$	$4,11 \cdot 10^4$	$1,21 \cdot 10^5$	$2,76 \cdot 10^5$
E_n	77,4	106	147	203	348	526
p_{E_n}	0,01%	0,01%	0,02%	0,03%	0,04%	0,07%

n	7	6	5	4	3	2
E_n^2	$7,40 \cdot 10^5$	$2,74 \cdot 10^6$	$6,05 \cdot 10^6$	$1,22 \cdot 10^7$	$3,02 \cdot 10^7$	$4,22 \cdot 10^8$
E_n	860	$1,66 \cdot 10^3$	$2,46 \cdot 10^3$	$3,49 \cdot 10^3$	$5,50 \cdot 10^3$	$2,05 \cdot 10^4$
p_{E_n}	0,11%	0,21%	0,31%	0,44%	0,70%	2,60%

n	1	0
E_n^2	$9,04 \cdot 10^8$	$6,25 \cdot 10^{11}$
E_n	$3,01 \cdot 10^4$	$7,90 \cdot 10^5$
p_{E_n}	3,81%	100%

Tabulka 1: Tabulka chyb E_n^2 , E_n a p_{E_n} pro sady měření M_{22}

n	5	4	3	2	1	0
E_n^2	$6,29 \cdot 10^5$	$1,29 \cdot 10^7$	$6,65 \cdot 10^7$	$1,55 \cdot 10^8$	$2,14 \cdot 10^9$	$2,70 \cdot 10^{12}$
E_n	793	3591	8156	$1,24 \cdot 10^4$	$4,63 \cdot 10^4$	$1,64 \cdot 10^6$
p_{E_n}	0,05%	0,22%	0,50%	0,76%	2,82%	100%

Tabulka 2: Tabulka chyb E_n^2 , E_n a p_{E_n} pro sady měření M_6

4 Výsledky aplikace analýzy hlavních komponent a jejich diskuse

Postup jsme použili na řadu chemických analýz roztoků odebraných z různých míst lokality Stráž pod Ralskem v různých časech. Pro ty účely jsme měli k dispozici dvě sady dat. První bylo 90 úplných analýz roztoků z různých míst lokality v různých časech s 22 složkami (označíme je M_{22}). Druhá sada dat obsahovala 638 analýz šesti hlavních složek roztoků (označíme ji M_6).

Na každé sadě měření jsme provedli analýzu hlavních komponent a vyčíslili jsme pro každé n chybu E_n^2 (1), která vznikne promítáním do optimálního podprostoru dimenze n . Výsledek je vyneseno v tabulkách 1 a 2. Zde E_n je odmocnina z E_n^2 a p_{E_n} je poměr chyby E_n a maximální chyby E_0 , což je odmocnina součtu druhých mocnin velikostí všech vektorů v množině M .

Z tabulek je patrné, že pro obě množiny M_{22} i M_6 je už při užití podprostoru s dimenzí 3 chyba aproximace nižší než 1%. Znamená to, že změřená data jsou výrazně korelovaná. Opačné zjištění by zamýšlený postup redukce dimenze znevěrohodňovalo.

Následovalo pozorování vlastností průmětů měření do zvoleného podprostoru z hlediska „přijatelnosti“ pro další zpracování. Pojem „přijatelnost“ nebyl nijak předem specifikován a bylo třeba ho definovat. Vzhledem k tomu, že každý vektor, se kterým pracujeme, musí být možno interpretovat jako potenciální chemickou analýzu roztoku, byla přirozeným požadavkem kladnost, přesněji nezápornost, každé složky průmětu. Dále musíme požadovat, aby průmět byl blízko k původnímu měření nejen v l^2 normě, ale také z hlediska každé složky, tj. v nějaké konkrétní vážené maximové normě $\|\mathbf{x} - \Pi_{V_n} \mathbf{x}\|_{\vec{\alpha}} = \max_i \alpha_i |x_i - (\Pi_{V_n} \mathbf{x})_i|$, kde $\vec{\alpha}$ vystihuje významnost každé

n	21	20	19	18	17	16
$\tilde{E}_n$	$9,62 \cdot 10^3$	$1,18 \cdot 10^4$	$1,87 \cdot 10^4$	$1,89 \cdot 10^4$	$1,95 \cdot 10^4$	$2,15 \cdot 10^4$
$p_{\tilde{E}_n}$	1,22%	1,49%	2,37%	2,39%	2,47%	2,72%
n	15	14	13	12	11	10
$\tilde{E}_n$	$2,62 \cdot 10^4$	$2,64 \cdot 10^4$	$2,78 \cdot 10^4$	$2,95 \cdot 10^4$	$3,72 \cdot 10^4$	$3,76 \cdot 10^4$
$p_{\tilde{E}_n}$	3,31%	3,33%	3,51%	3,73%	4,71%	4,76%
n	9	8	7	6	5	4
$\tilde{E}_n$	$3,77 \cdot 10^4$	$3,89 \cdot 10^4$	$3,91 \cdot 10^4$	$3,94 \cdot 10^4$	$4,67 \cdot 10^4$	$5,03 \cdot 10^4$
$p_{\tilde{E}_n}$	4,77%	4,92%	4,94%	4,98%	5,91%	6,36%
n	3	2	1	0		
$\tilde{E}_n$	$7,01 \cdot 10^4$	$7,51 \cdot 10^4$	$9,25 \cdot 10^4$	$7,90 \cdot 10^5$		
$p_{\tilde{E}_n}$	8,87%	9,50%	11,70%	100,00%		

Tabulka 3: Tabulka chyb $\tilde{E}_n$ a $p_{\tilde{E}_n}$ pro sady měření M_{22}

n	5	4	3	2	1	0
$\tilde{E}_n$	$4,72 \cdot 10^4$	$5,77 \cdot 10^4$	$8,33 \cdot 10^4$	$8,47 \cdot 10^4$	$9,17 \cdot 10^4$	$1,64 \cdot 10^6$
$p_{\tilde{E}_n}$	2,87%	3,51%	5,07%	5,16%	5,58%	100,00%

Tabulka 4: Tabulka chyb $\tilde{E}_n$ a $p_{\tilde{E}_n}$ pro sady měření M_6

složky. Jako přirozený výběr vektoru $\vec{\alpha}$ se nabízí vektor převrácených hodnot průměrů koncentrací jednotlivých látek v seznamu vzorků ($\alpha_i = 1/\bar{x}_i$), který seškáluje význam jednotlivých složek vzhledem k výrazně odlišným typickým koncentracím rozpuštěných iontů v roztocích.

Druhá část uvedené úvahy nás vedla k otestování modifikovaného postupu, ve kterém jsme kroky 2a a 8a modifikovaného algoritmu analýzy hlavních komponent nahradili následovně:

$$2b. \mathbf{X}^* = \mathbf{X} \cdot \mathbf{diag}(1/\bar{x}_1, \dots, 1/\bar{x}_s).$$

$$8b. \mathbf{Y} = \mathbf{Y}^* \cdot \mathbf{diag}(\bar{x}_1, \dots, \bar{x}_s).$$

Provádíme tím seškálování matice pozorování $\mathbf{X}$ vzhledem k průměrným hodnotám každé složky. Algoritmus tak nehledá lineární vektorový prostor minimalizující kvadratickou odchylku E_n (1) v l^2 normě, ale lineární vektorový prostor minimalizující kvadratickou odchylku

$$\tilde{E}_n^2 = \sum_{\mathbf{x} \in M} \|\mathbf{x} - \Pi_{V_n} \mathbf{x}\|_{2, \vec{\alpha}}^2. \quad (2)$$

ve škálované normě $\|\mathbf{x}\|_{2, \vec{\alpha}}^2 = \sum_{i=1}^n (\alpha_i x_i)^2$.

Vzhledem k podstatě algoritmu není možné ho snadno modifikovat pro optimalizaci přes nejvhodnější maximovou normu $\|\cdot\|_{\vec{\alpha}}$, proto jsme považovali tento postup za možnost přiblížení k dobrému výsledku, kterou je třeba prozkoumat.

Provedli jsme analýzu takto upraveným algoritmem pro stejné dvě množiny dat. Vektory průměrných hodnot zastoupení jednotlivých složek v roztoku byly pro tyto dvě množiny $\bar{\mathbf{x}}^{M_{22}} = (\bar{x}_1^{M_{22}}, \bar{x}_2^{M_{22}}, \dots, \bar{x}_{22}^{M_{22}})^T = (6,90 \cdot 10^4; 4,21 \cdot 10^4; 9,86 \cdot 10^2; 6,12 \cdot 10^3; 1,46 \cdot 10^3; 2,46 \cdot 10^2; 7,10 \cdot 10^1; 2,19 \cdot 10^1; 8,53 \cdot 10^{-1}; 2,97 \cdot 10^2; 5,59 \cdot 10^2; 2,62 \cdot 10^2; 1,30 \cdot 10^1; 2,53 \cdot 10^3; 5,78 \cdot 10^1; 2,18 \cdot 10^1; 3,74 \cdot 10^1; 1,78 \cdot 10^1; 6,10 \cdot 10^1; 1,25 \cdot 10^1; 6,33 \cdot 10^0; 1,57 \cdot 10^1)^T$ a $\bar{\mathbf{x}}^{M_6} = (\bar{x}_1^{M_6}, \bar{x}_2^{M_6}, \dots, \bar{x}_6^{M_6})^T = (4,69 \cdot 10^4; 2,84 \cdot 10^4; 6,68 \cdot 10^2; 4,15 \cdot 10^3; 1,04 \cdot 10^3; 1,63 \cdot 10^2)^T$.

dimenze	M_{22}		dimenze	M_6	
	neškálovaná	škálovaná		neškálovaná	škálovaná
1	0	0	1	0	0
2	6	6	2	0	2
3	2	7	3	0	7
4	7	4	4	3	0
5	6	5	5	0	1
6	7	7			
7	3	6			
8	7	1			
9	10	0			
10	5	0			
11	5	0			
12 až 21	0	0			

Tabulka 5: Počet průmětů prvků množiny M_{22} , resp. M_6 do optimálního podprostoru získaného jednou z modifikací analýzy hlavních komponent, které obsahují alespoň jednu zápornou složku

V tabulkách 3 a 4 jsou pro srovnání s výsledky v tabulkách 1 a 2 vyneseny kvadratické odchylky rekonstruovaných dat od původních dat v l^2 normě, což odpovídá l^2 normě rozdílu matic $\mathbf{X}$ a $\mathbf{Y}$ získané algoritmem s poslední popisovanou modifikací:

$$\tilde{E}_n^2 = \|\mathbf{X} - \mathbf{Y}\|^2. \quad (3)$$

Zřejmě došlo ke zvětšení odchylek měřených touto normou. Nemohlo tomu být jinak, pokud první modifikace algoritmu vedla k nalezení vždy optimálního podprostoru v neškálované l^2 normě.

Porovnali jsme výsledky těchto dvou postupů z hledisek námi definované „přijatelnosti“, tj. nezápornosti všech složek rekonstruovaných dat a malé vzdálenosti ve škálované maximové normě $\|\cdot\|_{\vec{\alpha}}$. První hledisko lze dobře vyhodnotit z tabulky 5, kde číslo vnesené pro každou dimenzi redukce označuje počet pozorování v původní množině M_{22} , resp. M_6 , jejichž průměty do podprostorů nalezených srovnávanými dvěma modifikacemi algoritmu analýzy hlavních komponent obsahují alespoň jednu zápornou složku.

Redukce na dimenzi 1 nevykazuje žádné složky se záporným průmětem, ani je vykazovat nemůže, protože promítáme vektory, které mají jen nezáporné složky do přímky, jejíž směrový vektor má samé nezáporné složky. Pokud chceme množinu M_{22} redukovat na podprostor s dimenzí větší než 1 „přijatelně“, musíme zvolit dimenzi větší než 11, nebo data škálovat a redukovat na dimenzi aspoň 9. Škálování tedy zlepšilo „přijatelnost“. Jistě by byla zajímavá analýza možnosti zlepšení tohoto aspektu „přijatelnosti“ jinou volbou škálovacího vektoru $\vec{\alpha}$. Tu jsme ale zatím neprovedli.

Množina M_6 naopak vykazuje zhoršení tohoto aspektu „přijatelnosti“ při použití škálování. „Přijatelnou“ redukci nelze provést při použití škálování pro žádnou dimenzi kromě 4. Naopak bez škálování je právě redukce na dimenzi 4 „nepřijatelná“.

Druhé hledisko „přijatelnosti“ redukce jsme zkoumali statistickým vyhodnocením maximové normy $\|\cdot\|_{\vec{\alpha}}$ rozdílu prvků původní množiny M_{22} , resp. M_6 a jejich průmětů do podprostoru nalezeného analýzou hlavních komponent. Pro ilustraci jsme vnesli do tabulky 6 vybrané statistické veličiny pro redukci množiny M_6 do dimenze 3 a 4 oběma modifikacemi analýzy hlavních komponent. Jsou zde uvedeny maximum, tj. největší odchylka jednotlivého měření od jeho obrazu v celé množině, aritmetický průměr odchylek přes celou množinu a medián odchylek přes množinu M_6 . Tabulka ukazuje, že škálování neredukuje vždy největší odchylku (pro dimenzi 3 maximum

	dimenze 3		dimenze 4	
	neškálovaná	škálovaná	neškálovaná	škálovaná
maximum	132	149	158	47
průměr	4,33	1,84	3,91	1,49
medián	0,18	0,08	0,17	0,04

Tabulka 6: Maximální a průměrná hodnota a medián maximové normy $\|\cdot\|_{\vec{\alpha}}$ rozdílu prvků množiny M_6 a jejich průmětů do podprostoru dimenze 3 a 4 získaného oběma variantami modifikace algoritmu analýzy hlavních komponent

dimenze	1.složka	2.složka	3.složka	4.složka	5.složka	6.složka
1	0	3	165	39	347	84
2	0	0	168	40	344	86
3	0	0	267	22	141	208
4	0	0	357	2	0	279
5	0	0	0	0	0	638

Tabulka 7: Frekvence výskytu maximálních chyb v jednotlivých složkách pro M_6 neškálovanou

škálováním vzrostlo). Pozorovali jsme ale typicky pokles průměru a mediánu. Významně vyšší průměr než medián naznačuje výskyt několika extrémně velkých odchylek. Analýza, zda největší odchylky vykazuje stejná, nebo různé podmnožiny množiny M_6 pro oba postupy redukce a pro různé dimenze nebyla zatím provedena.

Předpokládanou vlastností škálování je vyrovnaní významnosti jednotlivých složek vektorů měření. Tuto skutečnost lze dokumentovat tabulkami frekvencí výskytu maximální chyby v jednotlivých složkách měření 7, 8. Je zde vyneseno počet výskytů maximální odchylky v jednotlivých složkách vektorů z M_6 pro každou dimenzi redukce. Výskytem maximální odchylky v i -té složce vektoru $\mathbf{x} \in M_6$ myslíme rovnost $\alpha_i|x_i - (\Pi_{V_n}\mathbf{x})_i| = \|\mathbf{x} - \Pi_{V_n}\mathbf{x}\|_{\vec{\alpha}}$.

Z tabulek 7 je vidět, že postup bez škálování optimalizuje podprostor tak, že promítání generuje větší odchylky v procentuálním vyjádření ve složkách s menší průměrnou hodnotou $\bar{x}_i$. V tabulce 8 je patrná mnohem vyrovnanější distribuce chyby v důsledku škálování.

Z provedených analýz nevyplývá zatím jasný závěr o tom, který z uvedených postupů je vhodnější. Ukazuje se, že provedení analýzy způsobem, který je v předchozích odstavcích naznačen, je třeba udělat pro každou konkrétní úlohu podobného typu znovu. Vhodnost konkrétního zvoleného postupu pak závisí na konkrétní definici „přijatelnosti“ výsledků a dalších prioritách.

Pro naši aplikaci je součástí „přijatelnosti“ kladnost všech složek promítnutých dat. Tento požadavek nemusí být úplně striktní při případné aplikaci na jiný problém, kde bude možné část původních dat z analyzované množiny vyloučit nebo upravit (například omezením se na geometrickou nebo časovou podoblast modelované úlohy nebo přihlédnutím k možné chybě chemických

dimenze	1.složka	2.složka	3.složka	4.složka	5.složka	6.složka
1	0	14	142	59	315	108
2	23	73	336	109	25	72
3	76	171	2	243	17	129
4	163	235	1	239	0	0
5	592	0	1	45	0	0

Tabulka 8: Frekvence výskytu maximálních chyb v jednotlivých složkách pro M_6 škálovanou

analýz a podobně). Zajímavým směrem zkoumání může být konstrukce jiných škálovacích vektorů $\vec{\alpha}$ pro dosažení jiné distribuce významnosti jednotlivých složek přihlížející např. k různé přesnosti analýzy jednotlivých složek nebo k různé míře jejich významnosti (ve smyslu řídicích a ovlivněných složek roztoku).

Provedli jsme také porovnání analýzy hlavních komponent s postupem navrženým v odstavci 2. Výsledky srovnání jsou publikovány v ročníkovém projektu Ivana Bruského [1] a ukazují předpokládaný výsledek, že postup z odstavce 2 dává výsledky shodné s analýzou hlavních komponent, ale je pomalejší a jeho řešení obsahuje více technických problémů.

Chceme-li zde prezentovat další analýzu našich dat, musíme se omezit na dimenzi maximálně 3. Vzhledem k námi definovanému prvnímu požadavku „přijatelnosti“, nemáme veliký výběr a zvolíme pro další práci třírozměrnou redukci množiny M_6 získanou analýzou hlavních komponent bez škálování.

5 Aplikace inkrementální metody hledání konvexního obalu pro nalezení vhodných bázevých vektorů

Nalezení vhodné redukce původního prostoru popsaná v minulém odstavci je důležitým, ale ne posledním, krokem k simulaci transportu a chemických reakcí v mnohasložkovém roztoku s užitím redukované dimenze. Kromě identifikace vhodného podprostoru je třeba zvolit jeho vhodnou bázi, do které budeme rozkládat počáteční podmínky, ve které budeme provádět transportní výpočty a kterou budeme používat pro zpětnou rekonstrukci mnohasložkových dat.

Volba báze z hlediska matematiků není problém důležitý. Pokud si ale budeme vektory báze a souřadnice vektorů v ní interpretovat chemicky, zjistíme, že některé její vlastnosti významně pomůžou interpretaci konečného modelu.

Bázevých vektorů v mnohasložkovém prostoru lze interpretovat stejně jako všechny vektory reprezentující měření jako bázevé roztoky. Tedy bázevý vektor odpovídá v jistém smyslu chemické analýze nějakého bázevého roztoku. V takovém případě je ale na místě požadavek, aby všechny jeho složky byly nezáporné.

Souřadnice jednotlivých měřených roztoků v bázi redukovaného prostoru lze interpretovat jako směšovací poměry bázevých roztoků. Pak je ale na místě požadovat, aby tyto souřadnice byly nezáporné (a navíc jejich součet nebyl větší než jedna). Druhou podmínku píšeme do závorky, protože podaří-li se nám najít bázi s nezápornými složkami takovou, že všechna měření v ní mají nezáporné souřadnice, je splnění podmínky v závorce pouze věcí vhodného vynásobení všech bázevých vektorů kladným číslem.

Uvedené dva požadavky nejsou pro samotný model nutné, ale pro jeho interpretaci jsou významné a je-li možné najít takovou bázi, pak je přínosné ji mít k dispozici. Zřejmě je také to, že pokud taková báze existuje, rozhodně není jednoznačně určena. Na definici úlohy tak, aby byla jednoznačná, jsme se zatím nezaměřovali. Hledali jsme postup, jak zjistit, jestli taková báze vůbec existuje.

Za důležité považujeme zde poznamenat, že báze získaná analýzou hlavních komponent nemůže splňovat pro vyšší než první dimenzi už první z obou podmínek (nezápornost všech složek bázevých vektorů). První bázevý vektor sice všechny složky kladné mít bude zaručeně (směřuje v určitém smyslu směrem do středu množiny měření, která mají všechny složky nezáporné), další bázevé vektory jsou na něj ale kolmé, takže budou stejně zaručeně obsahovat záporné složky.

Aby zvolené vektory byly bázi určeného podprostoru V_n , která má výše zmíněné dvě vlastnosti, musí splňovat v geometrických pojmech následující podmínky:

- musí jich být tolik, kolik je dimenze zvoleného podprostoru V_n ,
- ležet ve zvoleném podprostoru V_n a být lineárně nezávislé,

- směřovat do prvního 2^s -antu prostoru V (mít jen nezáporné složky) a
- jimi uzavřená výseč podprostoru V_n musí obsahovat všechny průměty měření do podprostoru V_n (souřadnice průmětů měření v této bázi musí být nezáporné).

Pro představu poslední podmínky je třeba říct, že výsečí zde myslíme nekonečný jehlan příslušné dimenze s vrcholem v počátku, hranami ve směru jednotlivých báзовých vektorů a pláštěm tvořeným částmi rovin určených dvojicemi báзовých vektorů. Pro $n = 2$ jde o trojúhelníkovou výseč v rovině určenou dvěma vektory, pro $n = 3$ jde o trojboký jehlan určený třemi vektory atd.

První tři vlastnosti lze splnit velmi snadno a úloha postavená jen na nich bude mít zcela jistě nekonečně mnoho řešení pro každou úlohu, protože prostor V_n určený analýzou hlavních komponent obsahuje, jak jsme již řekli, vždy aspoň jeden vektor se samými kladnými složkami. Za první vektor nové báze bychom při řešení takové úlohy volili první báзовý vektor z analýzy hlavních komponent, za každý další i -tý vektor nové báze kombinaci prvního a i -tého báзовého vektoru z analýzy hlavních komponent obsahující nenulový násobek i -tého báзовého vektoru a mající nezáporné souřadnice. Poslední vlastnost úlohu technicky komplikuje, protože právě první báзовý vektor z analýzy hlavních komponent směřuje do středu měření a tak nemůže být součástí báze splňující také čtvrtou podmínku.

Pojem vhodné báze není, jak už jsme řekli, jednoznačný. Proto bude třeba pro konkrétní volbu vždy určit dodatečné požadavky. Volbu báze bez jednoznačných požadavků nemůžeme plně zalgoritmovat. Geometrická představa nám alespoň pomůže identifikovat postup, kterým zjednodušíme expertovi návrh vhodné báze provést. Navržený postup je následující:

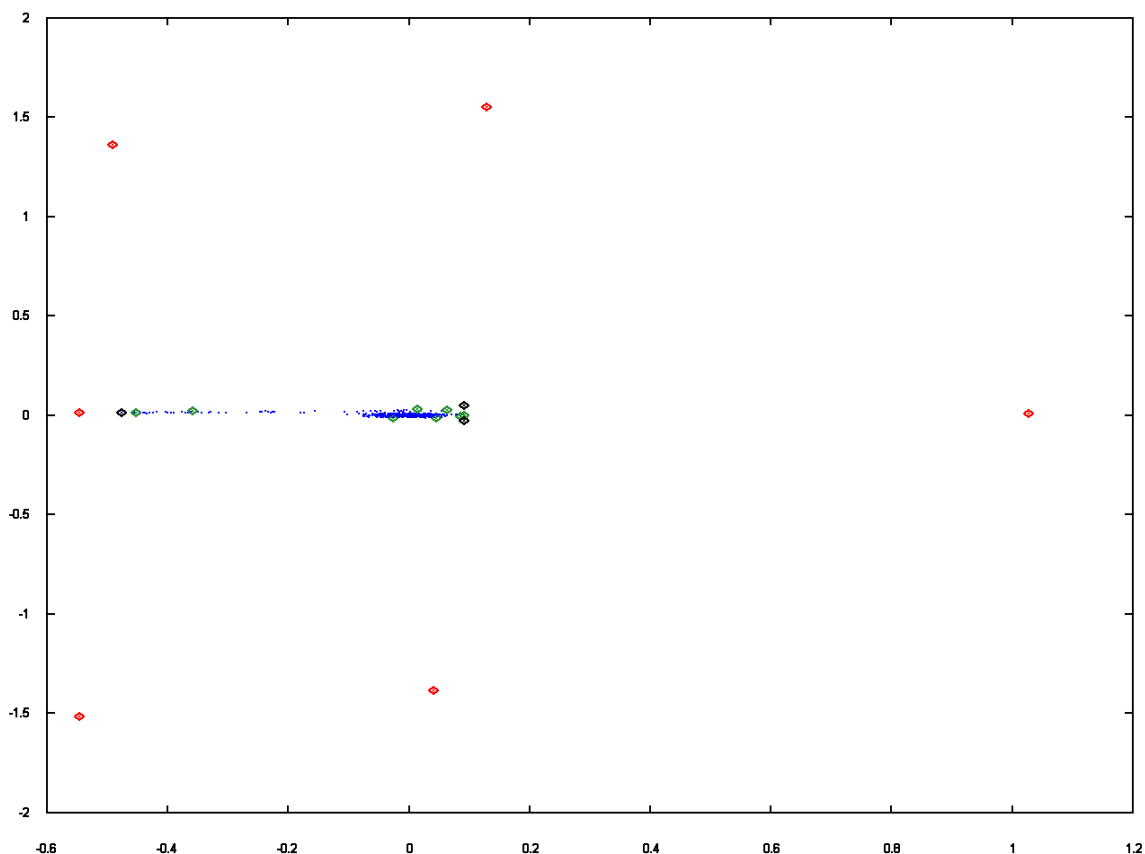
- provést pro množinu měření M v prostoru V analýzu hlavních komponent – zvolit podprostor V_n a průmět množiny M do tohoto prostoru $\tilde{M}$
- najít množinu průmětů vektorů standardní báze prostoru V do V_n (označíme ji $\tilde{E}$)
- promítnout množiny $\tilde{M}$ a $\tilde{E}$ v prostoru V_n na jednotkovou kouli se středem v počátku (dostaneme množiny $\tilde{\tilde{M}}$ a $\tilde{\tilde{E}}$ na sférické varietě $\tilde{V}_n$ s dimenzí $n - 1$)
- najít konvexní obal $\tilde{\tilde{M}}$ v $\tilde{V}_n$ (množinu jeho vrcholů označíme $\tilde{\tilde{M}}^-$) a konvexní obal $\tilde{\tilde{E}}$ v $\tilde{V}_n$ (množinu jeho vrcholů označíme $\tilde{\tilde{E}}^-$)

Poslední operaci navrženého postupu (hledání konvexního obalu dané množiny v prostoru obecné dimenze) můžeme realizovat například inkrementální metodou popsanou a demonstrovanou např. na webových stránkách Tima Lamberta z The University of New South Wales, Sydney [7].

Volba vhodné báze pak může proběhnout na $n - 1$ rozměrné varietě $\tilde{V}_n$. Vybereme množinu $\tilde{\tilde{B}}$ obsahující n konvexně nezávislých vektorů z $\tilde{V}_n$ tak, že konvexní obal $\tilde{\tilde{B}}$ v $\tilde{V}_n$ musí ležet celý uvnitř konvexního obalu $\tilde{\tilde{E}}^-$ a musí obsahovat celý konvexní obal $\tilde{\tilde{M}}^-$. Báзовé vektory pak navrhneme jako násobky vektorů z $\tilde{\tilde{B}}$ ve V_n takové, aby součet souřadnic každého prvku $\tilde{M}$ v nové bázi nebyl větší než 1.

Množina $\tilde{\tilde{B}}$ s uvedenými vlastnostmi nemusí pro obecně zvolenou množinu M a obecně zvolený prostor V_n existovat. Pokud existuje, nemusí být její volba jednoznačná.

Pro grafické přiblížení geometrického významu navrženého postupu jsme na obrázku 1 zobrazili výsledek navrženého postupu pro zvolenou sadu dat M_6 redukovanou do prostoru dimenze 3 analýzou hlavních komponent bez užití škálování. Z obrázku je patrné, že bázi lze v tomto případě volit velmi volně. Provedený výběr je proveden tak, že jeden z báзовých vektorů odpovídá přesně jednomu z měření (roztoku s nejmenším množstvím rozpuštěných látek v analyzované



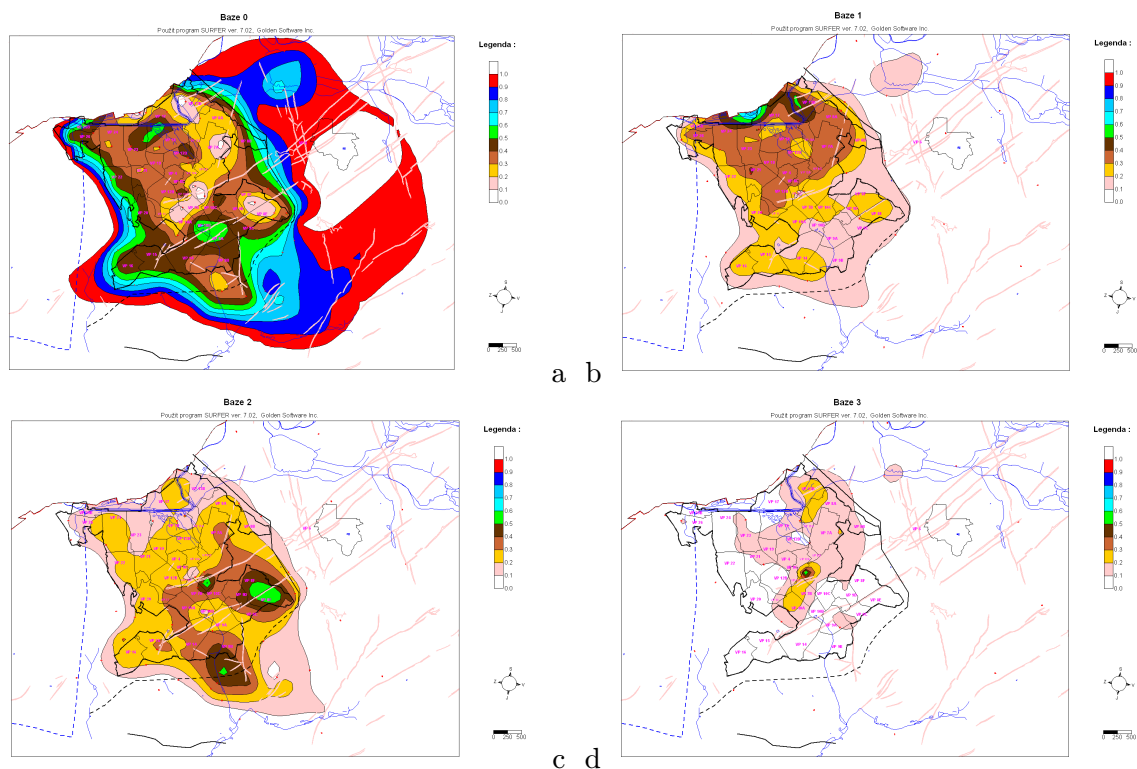
Obrázek 1: Průmět do jednotkové koule kolem počátku v podprostoru V_3 získaném PCA bez škálování. Modré body: množina $\tilde{M}_6$, červené zvýrazněné body: množina $\tilde{E}$, zelené zvýrazněné body: vrcholy konvexního obalu množiny $\tilde{M}_6$, černé body: zvolená báze optimálního 3D podprostoru $\tilde{B}$.

množině) a dva další jsou voleny co nejblíže jiným změřeným roztokům. Provedli jsme tak efektivní rozklad simulovaných roztoků do báze tří roztoků. Každý roztok v měřené množině tak lze interpretovat jako směs tří bazových roztoků (zastoupení každého z nich je dáno příslušnou souřadnicí v nové bázi) a destilované vody (jejíž chemický rozbor odpovídá nulovému vektoru a jejíž množství v každém měřeném roztoku je jedna mínus součet všech souřadnic v nové bázi).

6 Motivační příklad pro aplikaci postupu

V tomto odstavci budeme interpretovat možný význam navrženého postupu na výsledku Ing. Vladimíra Wasserbauera, CSc. z DIAMO, s. p. Ten v roce 2006 provedl rozklad chemických analýz provedených na vzorcích vod odebraných v části lokality Stráž pod Ralskem během jednoho roku do čtyř bazových roztoků. Jejich volba nebyla provedena navrženým postupem a není reprodukovatelná. Práce na návrhu tohoto postupu vyplynula z potřeby mít k dispozici postup použitelný pro analýzu větších množin dat bez apriorních znalostí použitých při návrhu této báze.

Prezentujeme zde výsledky rozkladu ing. Wasserbauera proto, že pro podobné grafické interpretace rozkladu množiny M_6 prezentovaného do minulého odstavce, nemáme dostatek informací o lokalizaci jednotlivých měření z množiny M_6 . Zde prezentované mapy byly vytvořeny ve s. p. DIAMO Ing. Jiřím Šrámkem s užitím programu SURFER a byly poskytnuty právě pro účel ilu-



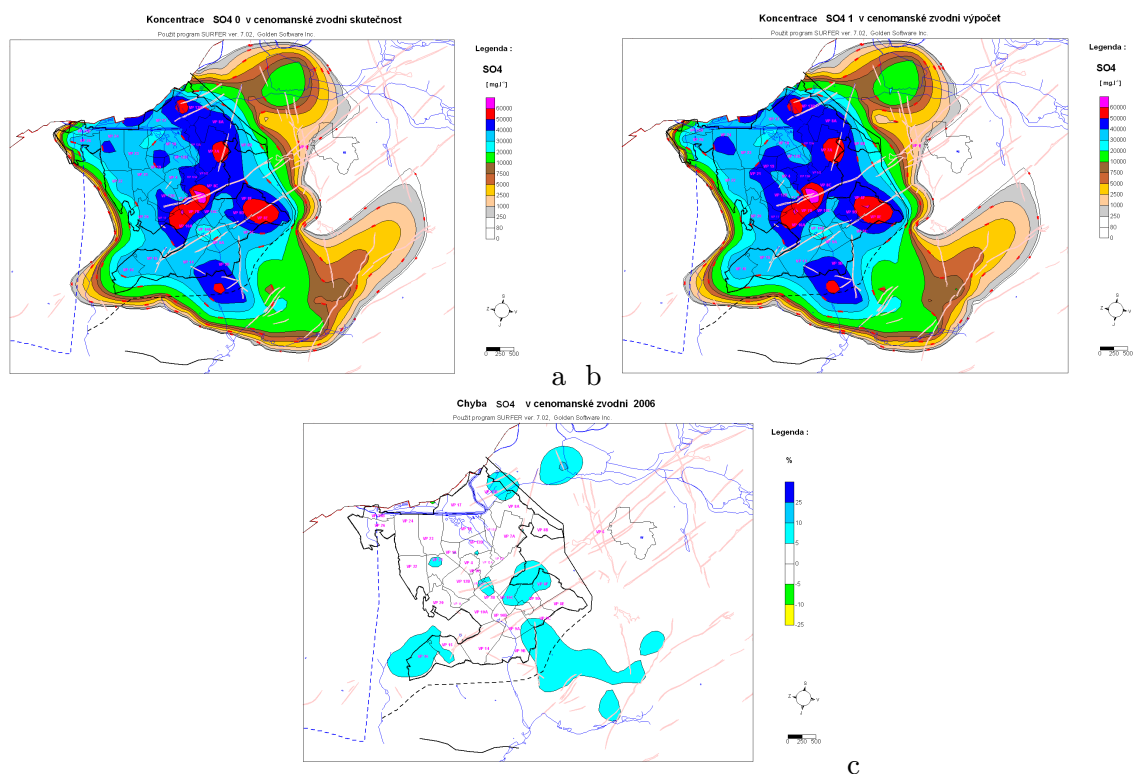
Obrázek 2: Souřadnice analyzovaných roztoků ve čtyřprvkové bázi vynesené do map. a) První bázeový vektor odpovídá složení neovlivněné cenomanské vody. b) Druhý bázeový vektor odpovídá složení technologických roztoků vtlačných v minulosti do severozápadní části lokality. c) Třetí bázeový vektor odpovídá složení technologických roztoků vtlačných v minulosti do jihovýchodní části lokality. d) Čtvrtý bázeový vektor odpovídá pravděpodobně produktům reakcí při mísení dvou druhů v minulosti vtlačných technologických roztoků.

strace možné interpretace výsledků navrhovaného postupu.

Analýzy chemických reakcí byly promítnuty do čtyřrozměrného podprostoru definovaného bázeovými vektory odpovídajícími chemickému složení tří skutečných roztoků z lokality a jednoho dalšího získaného optimalizací. První bázeový vektor odpovídá složení neovlivněné cenomanské vody, druhý a třetí bázeový vektor odpovídají složení technologických roztoků vtlačných v minulosti do dvou částí lokality. Čtvrtý vektor byl hledán tak, aby odchylky jednotlivých měření od jejich průmětu do čtyřrozměrného podprostoru byly minimální a neodpovídá žádnému skutečnému roztoku.

Na obrázku 2 je zobrazen rozklad analýz skutečných roztoků do výše popsané báze. Každý analyzovaný roztok byl odebrán z konkrétního místa v lokalitě a v mapě mu odpovídá příslušný bod. Programem SURFER byly hodnoty zobrazovaných veličin expandovány do plochy. Hodnoty vynesené na obrázku 2 odpovídají souřadnicím průmětů analyzovaných roztoků do čtyřrozměrného prostoru v uvedené bázi. Obrázky 2a až 2c ukazují dominantní zastoupení neovlivněné cenomanské vody v okolí vyluhovacích polí a dvou technologických roztoků v těch částech lokality, kde byly vtlačeny. Obrázek 2d ukazuje významnější zastoupení čtvrtého bázeového roztoku na rozhraní oblastí dominovaných technologickými roztoky a lze jej interpretovat tak, že dopočítaný čtvrtý roztok odpovídá reakčním produktům vznikajícím při mísení dvou technologických roztoků.

Obrázek 3a zobrazuje mapu koncentrace SO_4 podle provedených analýz roztoků. Na obrázku 3b je tatáž mapa provedená z dat promítnutých do čtyřrozměrného podprostoru. Obrázek 3c



Obrázek 3: a) Změřené koncentrace SO₄ vynesené do mapy. b) Koncentrace SO₄ zrekonstruované z dat redukovaných do čtyřrozměrného podprostoru vynesené do mapy. c) Rozdíly změřené koncentrace SO₄ a koncentrace SO₄ zrekonstruované z čtyřrozměrných dat vynesené do mapy.

ukazuje mapu rozdílu změřených a rekonstruovaných koncentrací SO₄.

Uvedený výsledek neodpovídá rozkladu do optimálního čtyřrozměrného podprostoru, přesto jsou pozorované rozdíly v obsahu SO₄ u většiny vzorků menší než 5% a u žádného vzorku nepřesahují 20%.

Reference

- [1] I. Bruský. Metodika snížení dimenze problému modelování transportu mnohasložkových vodných roztoků, 2007. Ročníkový projekt NTI TUL.
- [2] L. Zedek. Vyhledání a aplikace algoritmů informatiky vhodných pro snížení dimenze problému modelování transportu mnohasložkových vodných roztoků, 2007. Ročníkový projekt NTI TUL.
- [3] L. Svoboda. Conjoint analýza. *Acta Oeconomica Pragensia*, (1):193–204, 1995.
- [4] A. Černošek, V. Krajča, S. Petránek, and J. Mohylová. Praktické zkušenosti s aplikací metody analýzy nezávislých komponent a analýzy hlavních komponent pro eliminaci EEG artefaktů. *Lékař a technika*, (2), 2000.
- [5] K. Oborná. Analýza hlavních komponent indikátorů nezaměstnanosti v kontextu regionálního rozvoje. In *Sborník příspěvků z doktorandského semináře*, pages 489–496, Praha, 2005. ČZU PEF.

- [6] L. I. Smith. A tutorial on Principal Components Analysis, 2002. Studentská práce, University of Otago, Nový Zéland, 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal-components.pdf.
- [7] T. Lambert. Demonstrace algoritmů pro výpočet konvexního obalu. <http://www.cse.unsw.edu.au/~lambert/java/3d/>.