Vícerozměrná statistická analýza povodní na řece Sázavě v období 1961-2000

Milan Meloun, Jindřich Freisleben

Klíčová slova
PCA - FA - CA - CM - Cattelův graf - Shluková analýza - Dendrogram - Povodně - Sázava

Souhrn

Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních. Zdrojová matice dat obsahuje kvalitativní znaky v 13 sloupcích a sledované povodně jako objekty v 25 řádcích zdrojové matice dat. Cílem analýzy dat je nalézt shluk podobných si povodní a podobných znaků, jež povodně popisují. Podobnost povodní je posuzována na základě jistých podobností či vzdálenosti povodní v 13-rozměrném prostoru všech znaků dle kritéria, že čím je Mahalonobisova vzdálenost shluků povodní větší, tím menší je jejich vzájemná podobnost. Strukturu a vazby mezi sledovanými znaky vystihují metody snížení dimensionality. Rozptylový diagram skóre zobrazuje objekty, rozptýlené v rovině prvních dvou hlavních komponent (PCA) či faktorů (FA). Graf komponentních vah porovnává vzdálenosti (podobnosti) mezi znaky, kde krátká vzdálenost značí silnou korelaci dvou znaků. Znaky ale také povodně lze seskupovat do shluků *hierarchicky*, a to dle předem zvoleného způsobu metriky a výsledkem je *dendrogram*. Původních 13 sledovaných znaků lze zredukovat ve tři latentní proměnné, faktory nebo hlavní komponenty. Do nejvýznamnějšího faktoru FA1 se promítnou srážkové ukazatele a tání sněhu, druhý faktor FA2 souvisí se strmostí povodňové vlny a průtokem před povodní. Vysvětlený rozptyl v datech v prostoru tří nalezených faktorů je necelých 60 %, což znamená, že více než 40 % variability v datech bylo vyhodnoceno jako šum. Důvodem vysoké hodnoty šumu je jednak charakter dat, kdy povodňové stavy jsou sami o sobě extrémními případy, a jednak malý počet sledovaných povodní ku poměrně velkému počtu kvalitativních znaků. Optimum bývá totiž a nebo by mělo i zde být okolo 20 objektů na 1 sledovaný znak, a proto minimální poměr by měl být 5 povodní na 1 znak. Z hlediska analýzy objektů se povodně dělí zřetelně na dva významné shluky: povodně v zimním a povodně v letním období. Tyto významně odlišné skupiny by bylo ovšem optimální vyhodnocovat odděleně. V tom případě by ovšem poměr objektů a znaků byl však ještě nižší.

◆

1. Úvod

Povodňový stav je odrazem souběhu hydrologických a meteorologických událostí, které lze kvantifikovat hodnotami sledovaných parametrů. Základní otázka, na kterou hledáme odpověď, je jak s dostatečným předstihem předpovědět povodeň. Prvním krokem je ve vytypovaných dlouhodobě sledovaných parametrech určit ty znaky, jejichž hodnoty odráží existenci povodně a míru jejího dopadu. Usilujeme proto o nalezení statisticky významných znaků, přispívajících k rozlišení mezi povodňovými stavy. Právě pro tento účel je vhodné užití metod s latentními proměnnými k určení vnitřní struktury dat. Na řece Sázavě bylo v období 1961-2000 monitorováno 25 povodňových událostí, označených zde jako případů, které byly popsány hodnotami 13 sledovaných znaků. Cílem bylo nalezení vzájemných vztahů mezi znaky a snížení rozměrnosti popisu čili nalezení menšího a postačujícího počtu faktorů.

2. Data

Zdrojová matice rozměru 25 × 13 obsahuje následující znaky ve sloupcích: **Qkul** značí kulminační průtok čili maximální průtok během povodně, **Qo** je průtok v patě vlny čili průtok těsně před započetím rostoucího trendu na křivce průtoků za čas, **INTvz** je intenzita vzestupu nebo-li změna průtoku z **Qo** na **Qkul.** za čas, **INTp** je intenzita poklesu čili změna průtoku z **Qkul.** zpět na **Qo** za čas, **Thy** je vzdálenost příčinné srážky od vrcholu povodňové vlny na časové ose, **Psr** je množství příčinné srážky na oblast příslušného povodí, **Tpsr** je trvání příčinné srážky, **MAXsr** je maximální denní srážka v průběhu příčinné srážky, **UPS** je ukazatel předchozích srážek za 30 dní před kulminací povodně, kterým se vyjadřuje nasycenost povodí, **Dmaxsr** je počet dní mezi výskytem maximální srážky a kulminací, **Pvrch** je počet vrcholů povodňové vlny nebo-li počet maxim na křivce vyjadřující závislost průtoku na čase, **DT** je změna teploty v letním období ochlazení a v zimním oteplení, **US** je úbytek sněhu. V řádcích je uváděno i datum, kdy se povodeň na řece Sázavě vyskytla a rovněž řádkový index tohoto data –viz **Tab. 1.**

3. Exploratorní analýza

Protože hodnoty znaků jsou vyjádřeny v různých jednotkách a rozličném měřítku, je nutné před vlastní analýzou provést jejich standardizaci. Byla proto zvolena studentizace (neboli *t*-transformace) dle vzorců

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

, kde $\bar{x}_j$ je sloupcový průměr a s_j je sloupcová směrodatná odchylka. Studentizace vede k vycentrování znaků, kdy průměr prvků ve sloupci se rovná nule a rozptyl je roven jedné. Při

Tabulka 1. Zdrojová matice dat povodní na řece Sázavě

Index	Povodňová vlna	Qkul. [m³.s⁻¹]	Qo [m³.s⁻¹]	INTvz [m³.s⁻¹.h⁻¹]	INTp [m³.s⁻¹.h⁻¹]	Thy [h]	Psr [mm]	Tpsr [den]	MAXsr [mm]	UPS [mm]	Dmaxsr [den]	Pvrch [počet]	DT [C]	US [cm]
1	11.2.1961	204	33	3.6	1.1	40.2	27.8	3	14.3	28.2	2	2	4.9	6.5
2	6.6.1961	138	31.8	1.3	0.4	63.4	55	5	19.3	69.4	3	4	5.5	0
3	1.4.1962	161	68.2	2.2	0.7	50.7	10.3	3	4	15.8	3	2	6.1	2.2
4	15.5.1962	264	44.6	6.1	2.2	53.2	47.7	3	24.2	67.8	2	1	5.4	0
5	17.6.1962	152	16.3	10.4	1.5	32.7	18.2	2	14.2	29.2	1	1	2.6	0
6	20.3.1965	257	37.9	3	0.9	88.6	2.4	1	2.4	11.7	4	3	3.4	32.4
7	12.5.1965	202	63.9	3.3	1.9	63.5	45.4	5	22.7	60.4	2	1	5	0
8	7.6.1965	250	84	2.7	0.9	57.9	36.2	3	16.1	74.2	3	1	1.4	0
9	19.7.1965	381	37.9	9.5	5.2	49.1	76.8	4	56.1	80.8	2	2	6	0
10	10.2.1966	176	79.8	2.7	1.2	60.2	20.6	3	11.1	27.5	2	1	3	1.3
11	28.8.1966	196	52	3.9	0.9	51.5	34.6	2	25.2	69.8	2	2	4.2	0
12	4.2.1967	223	51	4.2	1.1	56.2	10.7	2	8.3	19.9	2	1	7.6	7.6
13	25.12.1967	234	20.4	4.5	2.9	61.5	30.6	3	18.4	40.4	2	1	12.2	24.7
14	15.3.1969	208	34.9	2.1	1.8	44.9	20.7	4	11	12.7	1	2	7.6	7.8
15	26.3.1970	245	41.7	1.6	1.8	95.2	3.3	2	1.8	16.7	4	2	4	25.9
16	11.4.1970	139	73.4	1.9	0.9	39.2	13	2	11.2	17.3	2	1	1.9	1.6
17	21.5.1972	170	28.2	3.4	1.4	59.1	64.2	4	41.8	65	2	1	2.8	0
18	9.12.1974	228	64.8	3.4	1	64.9	53.9	5	17.2	59.4	2	2	6.3	3.8
19	3.7.1975	142	34.2	2.8	1.3	64.7	55.3	4	23.1	79.2	2	1	2.4	0
20	16.1.1976	166	33.3	1.2	0.9	74.7	33.4	5	12.5	47.1	5	2	2	1.2
21	21.7.1981	158	9.5	3.5	1.5	42.7	98.4	4	45.5	106.6	2	1	8.4	0
22	19.3.1993	134	20.4	1.5	1.1	73	4.8	2	2.8	13.3	2	1	5.7	14.3
23	16.9.1995	135	13.8	2.8	1.6	45.8	46.2	2	27.8	79.1	2	1	8	0
24	4.3.1999	170	56.3	2.5	0.6	49.8	8.5	2	5.6	27.8	2	2	5	19.3
25	31.3.2000	145	95	0.7	0.6	57.6	27.4	4	12.7	52.3	2	1	4.3	0

Tabulka 2. Základní popisné statistické charakteristiky znaků polohy, rozptýlení a tvaru rozdělení

		Qkul.	Qo	INTvz	INTp	Thy	Psr	Tpsr	MAXsr	UPS	Dmaxsr	Pvrch	DT	US
Původní data	Objektů	25	25	25	25	25	25	25	25	25	25	25	25	25
	Průměr	195.1	45.1	3.4	1.4	57.6	33.8	3.2	18.0	46.9	2.3	1.6	5.0	5.9
	Rozptyl	3260.9	536.4	5.3	0.9	213.2	597.7	1.4	184.5	756.6	0.8	0.6	6.2	92.1
	Medián	176.0	37.9	2.8	1.1	57.6	30.6	3.0	14.3	47.1	2.0	1.0	5.0	1.2
	Spod. kvartil	152.0	31.8	2.1	0.9	49.1	13.0	2.0	11.0	19.9	2.0	1.0	3.0	0.0
	Horní kvartil	228.0	63.9	3.6	1.6	63.5	47.7	4.0	23.1	69.4	2.0	2.0	6.1	7.6
	Šikmost	1.5	0.5	2.0	2.8	0.8	0.8	0.2	1.3	0.3	1.5	1.6	0.9	1.7
	Špičatost	3.2	-0.6	4.2	9.7	1.0	0.5	-1.1	1.8	-1.0	2.7	3.0	1.3	1.8
	Minimum	134.0	9.5	0.7	0.4	32.7	2.4	1.0	1.8	11.7	1.0	1.0	1.4	0.0
	Maximum	381.0	95.0	10.4	5.2	95.2	98.4	5.0	56.1	106.6	5.0	4.0	12.2	32.4
Data po transformaci a standardizaci	Průměr	0	0	0	0	0	0	0	0	0	0	0	0	0
	Rozptyl	1	1	1	1	1	1	1	1	1	1	1	1	1
	Medián	-0.3	-0.2	0.0	-0.1	0.7	0.0	-0.1	0.0	0.1	-0.2	-0.9	0.1	0.7
	Spod. kvartil	-0.8	-0.5	-0.5	-0.5	-1.4	-0.8	-1.0	-0.3	-1.0	-0.2	-0.9	-0.8	-1.0
	Horní kvartil	0.7	0.9	0.4	0.6	0.7	0.7	0.7	0.6	0.9	-0.2	1.1	0.6	1.0
	Šikmost	1.1	0.0	-0.3	0.0	-0.8	0.1	0.0	-0.1	0.0	-0.1	0.3	0.1	-0.1
	Špičatost	1.5	-0.7	0.9	0.8	-1.4	-0.6	-0.9	-0.2	-1.4	1.7	-2.1	-0.2	-2.1
	Minimum	-1.2	-1.9	-2.5	-2.4	-1.4	-1.7	-2.1	-1.9	-1.5	-2.3	-0.9	-1.9	-1.0
	Maximum	2.9	1.9	2.0	2.3	0.7	2.1	1.5	2.0	1.8	2.2	1.2	2.3	1.1

používání metod k určení vnitřní struktury ve znacích jako je metoda hlavních komponent PCA a faktorová analýza FA vycházíme z předpokladů o datech, jako je neexistence odlehlých bodů, vícerozměrná normalita atd. Předpoklady je třeba vyšetřit a ověřit.

Krabicový graf (**Obr. 1.**), jako jedna z důležitějších diagnostik exploratorní analýzy jednorozměrných dat, ukazuje mírné zešikmení u většiny znaků. Extrémním případem jsou však znaky **INTvz, INTp, Tpsr, Dmaxsr, Pvrch** a **DT**, které se oproti ostatním vyznačují velmi nízkou proměnlivostí, a tím pádem málo přispívají k rozlišení mezi znaky. V rozdělení dat sloupců zdrojové matice je patrný silný logaritmický trend, a bývá zvykem v rámci předúpravy dat provést jejich transformaci.

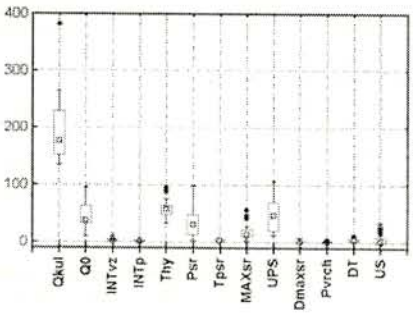
Symbolové (ikonové) grafy na **obr. 2.** představují důležité diagnostiky exploratorní analýzy vícerozměrných dat a umožňují porovnávat podobnost jednotlivých povodní, a to na základě podobnosti grafických obrazců. Sledované povodně se na hvězdicovém grafu jeví poměrně odlišně. Pokud slevíme z nároků na dokonalou shodnost obrazců, je možné si z určité podobnosti hvězdiček povšimnout, například podobnost povodní v bloku (3-24-10-16), nebo podobnost i v jiných blocích povodní (1-14), (2-19-23), (4-17-19), atd, kde pořadová čísla povodně se týkají data, uskutečněné povodně – viz **tab. 2.**

Porovnáním základních charakteristik znaků zdrojové matice bez předúpravy a s předúpravou dat transformací ukazují, že po transformaci je patrný posun k normálnímu rozdělení. Hodnoty mediánu a aritmetického průměru se potom k sobě blíží až se téměř rovnají, dále hodnota šikmosti se blíží k nule, a absolutní hodnoty rozdílů středních hodnot od horního a dolního kvartilu jsou si také blízké. Po standardizační předúpravě dat se hodnoty aritmetických průměrů rovnají nule a hodnoty rozptylů jedné. Transformace dat by vedla rovněž ke snížení rozdílu extrémů (maxim a minim) od středních hodnot.

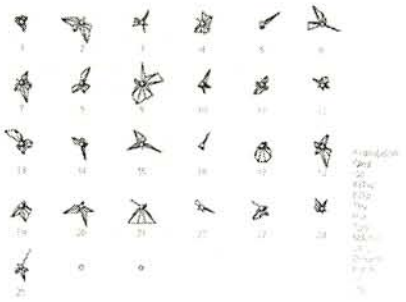
Hodnoty korelačních koeficientů v **tab. 3.** jsou první informací o vzájemných vazbách mezi znaky. První vzájemný vztah je patrný u znaků **Qkul, INTvz** a **INTp**. Tyto znaky totiž úzce souvisejí s tvarem průtokové křivky. Na druhém vztahu se podílejí znaky **Tpsr, Psr, MAXsr** a **UPS**, t.j. vesměs znaky charakterizující srážky. Zajímavý je také vztah úbytku sněhu **US** ke znakům ve druhé skupině, který je s nimi v negativní korelaci. To odpovídá skutečnosti, že i nižší srážky ale ve spojení s táním sněhu mohou rovněž vyvolat povodeň.

Tabulka 3. Korelační matice znaků. Tučně značené korelace jsou významné na hladině p < 0.05.

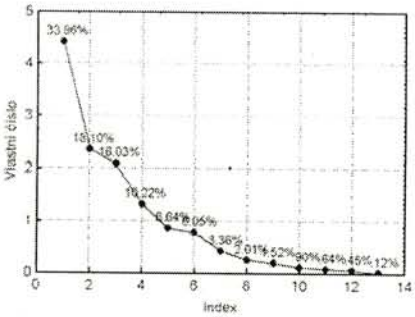
	Qkul.	Qo	INTvz	INTp	Thy	Psr	Tpsr	MAXsr	UPS	Pvrch	DT	US
Qkul.	1											
Qo	0.14	1										
INTvz	**0.51**	-0.31	1									
INTp	**0.56**	-0.37	**0.63**	1								
Thy	0.05	0.35	-0.30	-0.23	1							
Psr	0.08	-0.24	0.28	0.28	-0.12	1						
Tpsr	-0.02	0.08	-0.18	0.02	0.16	**0.70**	1					
MAXsr	0.13	-0.25	**0.44**	0.39	-0.24	**0.94**	**0.53**	1				
UPS	0.05	-0.16	0.21	0.16	0.01	**0.92**	**0.53**	**0.86**	1			
Pvrch	0.30	0.06	-0.12	-0.25	-0.08	-0.18	0.06	-0.27	-0.23	1		
DT	0.15	-0.38	0.18	0.30	-0.15	0.14	0.06	0.12	0.03	0.06	1	



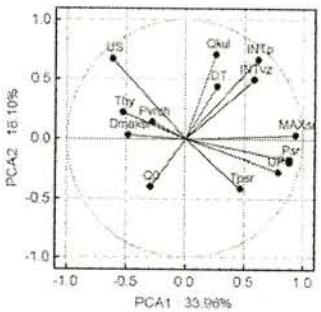
Obr. 1. Krabicový graf proměnlivosti znaků u 25 povodní na řece Sázavě



Obr. 2. Hvězdicový graf ukazuje na podobné povodně



Obr. 3. Cattelův indexový graf úpatí vlastních čísel



Obr. 4. Graf komponentních vah 1. a 2. hlavní komponenty

4. Metoda hlavních komponent

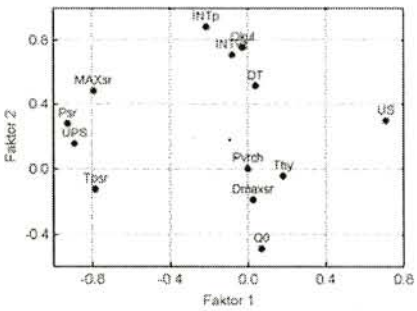
Cattelův indexový graf úpatí vlastních čísel na **obr. 3.** představuje základní pomůcku k určení potřebného počtu hlavních komponent u metody hlavních komponent PCA. Indikuje zde totiž prvním výraznějším zlomem u třetího bodu první 3 hlavní komponenty. Použijeme-li však Kaiserovo kritérium 1, pak nad hodnotou 1 leží první 4 hlavní komponenty, které ovšem nelze graficky zobrazit. Zůstaneme proto u prvních tří hlavních komponent, které vysvětlují celkový rozptyl pouze ze 67,63%. Pokusme se vysvětlit polohu znaků v grafu komponentních vah na **obr. 4.**

1. hlavní komponenta (PCA1) na **obr. 4.** se týká šesti významných

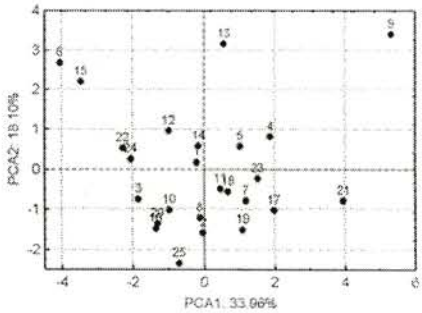
a poměrně faktorově čistých znaků, které mají vůči PCA1 vysokou hodnotu komponentní váhy a jsou to **INTp, INTvz, MAXsr, Psr, UPS,** a **Tpsr.** Záporná hodnota znaku **US** znamená, že hodnota znaku roste s klesající hodnotou PCA1. PCA1 tedy roste s množstvím srážek a klesá s úbytkem sněhu.

2. hlavní komponenta (PCA2): nejvýznamnější komponentní váhu mají čtyři znaky **Qkul, US, INTvz** a **INTp**, klesá se strmostí povodňové vlny.

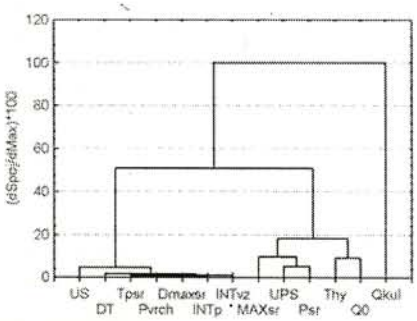
Obecně lze formulovat závěry, že graf komponentních vah pro PCA1 a PCA2 na **obr. 4.** ukazuje, jak souřadnice každého znaku vyjadřují svůj příspěvek dotyčného znaku do hlavní komponenty. Přičemž platí, že důležité znaky leží daleko od počátku a méně důležité blízko počátku. Malý úhel mezi dvěma průvodiči znaku vyjadřuje silnou pozitivní korelaci obou znaků, zatímco úhel okolo 180° vyjadřuje silnou negativní korelaci. Je-li úhel mezi průvodiči znaků okolo 90°, jsou oba znaky vzájemně naprosto nekorelované. Dále platí pravidlo, že znaky v tomto grafu blízko sebe jsou si vzájemně podobné, zatímco znaky daleko od sebe jsou si nepodobné. (1) Znaky **MAXsr, Psr, UP** a **Tpsr** jsou středně až silně pozitivně korelované a přispívají především do PCA1. (2) Znak **US** přispívá negativně zejména do PCA1 a je přitom se znaky v 1. skupině v silné negativní korelaci. (3) Znaky **INTvz** a **INTp** ale také **Qkul** a **DT** jsou silně pozitivně korelované a přispívají zejména do PCA2, přičemž se znaky v předchozích dvou skupinách téměř nekorelují, protože s nimi vykazují téměř pravý úhel.



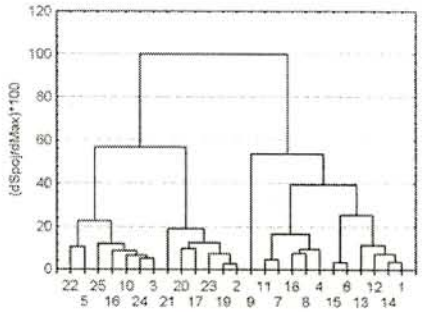
Obr. 5. Graf faktorových vah po Varimax rotaci pro 1. a 2. faktor



Obr. 6. Graf faktorového skóre po Varimax rotaci pro 1. a 2. faktor ukazuje na shluky podobných povodní



Obr. 7. Dendrogram znaků povodní odhaluje podobnostní shluky



Obr. 8. Dendrogram povodní odhaluje povodně podobné a blízké si charakterem všech sledovaných znaků

5. Faktorová analýza

Pohledem na graf faktorových vah na **obr. 5.** je patrné, že mírným pootočením faktorových os bychom docílili lepšího proložení skupin významnějších znaků. Proto nyní použijeme metodu faktorové analýzy, která zmíněnou potřebnou rotaci faktorů umožní. Faktorová analýza je metoda, jejíž výsledky jsou závislé na předem zvoleném optimálním počtu latentních faktorů. K odhadu vhodného počtu faktorů je opět používán Cattelův indexový graf úpatí vlastních čísel, nicméně některé softwary uvádějí i jiné způsoby, jako je například test významnosti přidání další latentní proměnné. K určení počtu latentních proměnných (faktorů) je rozhodující porovnání hodnot vypočítané experimentální hodnoty F a tabelární hodnoty $F(0,95)$ pro 95% statistickou jistotu. Pokud je experimentální F vyšší než kritická hodnota $F(0,95)$, je přídání další latentní proměnné statisticky významné. V našem případě se ukázalo, že dle tohoto kritéria je potřebný počet faktorů roven 2. Pro rozhodnutí o počtu latentních proměnných se obvykle dává přednost hodnotě určené z Cattelova indexového grafu úpatí vlastních čísel, a proto provedeme faktorovou analýzu raději pro 3 faktory s vysvětleným rozptylem 58.60 %. **Obr. 5.** přináší graf faktorových vah po rotaci Varimax pro FA1 a FA2. Faktor 1 (FA1) ukazuje na tři významné faktorově čisté znaky **Psr, MAXsr, UPS** s kladnou hodnotou faktorové váhy, zatímco zápornou hodnotu faktorové váhy mají znaky **US, Thy, Dmaxsr, Pvrch**, ale také **Qo**, což znamená, že hodnota znaku roste s klesající hodnotou FA1. Tedy FA1 roste s množstvím srážek a klesá s úbytkem sněhu. U faktoru 2 (FA2) mají nejvýznamnější kladnou faktorovou váhu znaky **INTvz, INTp, Qkul** a **DT**. FA2 roste se strmostí povodňové vlny a klesá s výší průtoku před povodní (záporná faktorová váha u znaku **Qo**). U grafu faktorových vah lze na základě polohy zobrazených znaků vyvozovat obdobné závěry o jejich vzájemné korelaci jako u grafu komponentní vah metody hlavních komponent. To, že se poloha znaků v grafu 6 a 7 neshoduje, je důsledkem provedené rotace faktorů v rámci faktorové analýzy. Snahou bylo docílit přiblížení os nejvýznamnějším sledovaným znakům, a tak docílit většího počtu faktorově čistých znaků, což značí co nejvyšší hodnotu faktorové zátěže u jednoho z faktorů a naopak co nejnižší u ostatních.

(a) Znaky **MAXsr, Psr, UPS** a **Tpsr** na **obr. 5.** jsou středně až silně vzájemně pozitivně korelované, přispívají záporně zejména do FA1. Dále **US** přispívá kladně také do FA1, ale se znaky v 1. skupině je v silné negativní korelaci. Znaky **INTvz, INTp, Qkul** a **DT** jsou vzájemně silně pozitivně korelované, přispívají kladně do FA2, a se znaky v předchozích dvou skupinách téměř nekorelují. Znak **Qo** a **Dmaxsr** přispívá záporně do FA2, a s ostatními znaky ve FA2 je v silné negativní korelaci. V porovnání s metodou hlavních komponent došlo po faktorové analýze k posílení „faktorové čistoty" znaků, zvětšil se rozdíl mezi faktorovými vahami. Znak **Qo** značící průtok v patě vlny se stal významným pro FA2. Došlo k potvrzení závěrů z PCA. Důležitým cílem faktorové analýzy je výstižné pojmenování nově zavedených faktorů: pro FA1 je to pojmenování *Poměr původu vody (srážky ku sněhu)*, pro FA2 pojmenování *Poměr strmosti vlny ku průtoku před povodní.*

Na **obr. 6.** je graf komponentního (nebo také faktorového) skóre u PCA nebo FA po varimax rotaci PCA (nebo také FA) pro první dvě hlavní komponenty PCA1 a PCA2, kde objekty ve shluku jsou si podobné a nepodobné objektům v ostatních shlucích, osamělé objekty jsou podezřelé z odlehlosti a platí pravidlo, že ideální je rozmístění objektů po celé ploše diagramu. V grafu obr. 6. PCA1 roste s množstvím srážek a klesá s úbytkem sněhu a PCA2 klesá se strmostí povodňové vlny. Svislá linie dělí oblast objektů na povodně v zimním a letním období. Pro „zimní" povodně platí nižší hodnoty PCA1 postačují nižší srážky vlivem významného podílu vody z tání sněhu. Naopak pro letní povodně, jsou typické vyšší hodnoty PCA1 na vzniku povodně se podílí větší množství srážek, protože logicky na průtoku vody se nepodílí tající sníh. PCA2 nepřispívá tak významně k rozlišení povodní v zimním a letním období. Je možné si pouze povšimnout, že o trochu větší počet „letních" povodní nabývá hodnot více odlehlých od středu, tedy pro „zimní" povodně platí větší stabilita hodnot znaků INTvz a INTp. To je patrně z důvodu, že v zimním období se podíl vody z tajícího sněhu dostává do toku pozvolněji než voda z intenzivních srážek v letním období. V grafu je možné identifikovat několik povodňových událostí, které se odlišují od skupiny ostatních kam podle období průběhu náleží:

a) U obou povodní 18 (ze dne 9.12.1974) a 13 (ze dne 25.12.1967) sehrály srážkové faktory významnější roli než bývá u zimních povodní obvyklé.
b) U ojedinělé povodně 25 (ze dne 31.3.2000) je poměrně vysoká hodnota srážek a téměř žádný příspěvek vody z tajícího sněhu.
c) U shluku dvou povodní 15 (ze dne 26.3.1970) a 6 (ze dne 20.3.1965) jsou nižší hodnoty srážek s vysokými úbytky tajícího sněhu.
d) U povodně 5 (ze dne 17.6.1962) je extrémně vysoká hodnota intenzity vzestupu povodňové vlny, kratší doba mezi příčinnou srážkou a kulminací, ostatní sledované znaky spíše podprůměrné v porovnání s ostatními povodněmi.
e) U odlehlého bodu - povodně 9 (ze dne 19.7.1965) existuje obdobně strmý růst povodňové vlny jako u povodně 5 (ze dne 17.6.1962), ovšem s dalece intenzivnějšími srážkami, a to i v období před povodní (vysoká hodnota **UPS**) a extrémně vysoká hodnota kulminačního průtoku.
f) U okrajového bodu grafu - povodně 21 (ze dne 21.7.1981) jsou extrémní hodnoty srážkových ukazatelů, zatímco ostatní jsou nižší.
g) U povodně 2 (ze dne 6.6.1961) jsou nižší hodnoty vzestupu a poklesu povodňové vlny, které jsou typičtější spíše pro povodně v zimním období.

6. Shluková analýza

Pro analýzu shluků byla použita jako míra vzdálenosti Euklidovská vzdálenost a za metodu shlukování pak Wardova metoda. Volba míry vzdálenosti je odvislá od míry korelace mezi znaky. Pro silně korelované znaky je vhodnější použít Mahalanobisovu vzdálenost. Volba vhodného aglomeračního způsobu je v některých softwarech usnadněna možností výpočtu *kofenetického korelačního koeficientu*, kdy platí pravidlo, že čím je jeho hodnota blíže 1, tím je tato metoda vhodnější. Jiné kritérium je kritérium *delta*, které indikuji za nejvhodnější metodu tu, která má hodnotu nejbližší nule. Protože metoda analýzy shluků nerozlišuje významné proměnné od nevýznamných, zvyšujících šanci odlehlých bodů, závisí na volbě proměnných nalezení správných shluků. Dendrogram znaků na **obr. 7.** potvrzuje závěry o znacích z dosavadní provedené vícerozměrné analýzy:

1. shluk obsahuje jednak znaky **UPS, MAXsr** a **Psr** a k nim se připojují také dva znaky **Thy** a **Qo**, které popisují množství srážek a jsou si nejpodobnější, a dále nejvíce pak maximální denní srážka a příčinná srážka.

2. shluk obsahuje znaky **DT, Tpsr, Pvrch, Dmaxsr, INTp, INTvz**, které popisují tvar povodňové vlny a k nim se pojí i znak **US**.

Zbývající znak **Qkul**, který, jak víme, je k ostatním buď v silné negativní korelaci nebo s nimi nekoreluje vůbec a je k výše uvedeným znakům nejméně podobný.

Dendrogram povodní na **obr. 8.** vede k těmto závěrům:

1. shluk obsahuje povodně 22, 5, 25, 16, 10, 24 a 3. Znaky **Qkul, Psr** a **MAXsr** vykazují nízké až podprůměrné hodnoty, znak **UPS** pak nízké až podprůměrné hodnoty, s výjimkou povodně 25 mající nadprůměrnou hodnotu. Znaky **INTvz** a **INTp** mají nízké až podprůměrné hodnoty, vyjma povodně 5 s extrémní hodnotou u **INTvz**, zatímco znak **Qo** má nadprůměrné až velmi vysoké hodnoty s výjimkou povodní 5 a 22 s hodnotami nízkými a znak **US** vykazuje nulové až nízké hodnoty, vyjma povodní 22 a 24 s vysokými hodnotami.

2.shluk obsahuje povodně 21, 20, 17, 23, 19, 2. Znaky **Qkul, Qo** zde vykazují velmi nízké až podprůměrné hodnoty, znaky **Psr** a **UPS** mají nadprůměrné až velmi vysoké hodnoty, znak **MAXsr** vysoké hodnoty s výjimkou povodně 20 s podprůměrnou hodnotou a znak **US** nulové hodnoty, kde povodeň 20 má nízkou hodnotu.

3.shluk obsahuje povodně 11, 7, 18, 8, 4. Znaky **Qkul, Qo, UPS** a **INTvz** mají nadprůměrné až vysoké hodnoty, znak **MAXsr** průměrné až vysoké hodnoty, znak **Psr** nadprůměrné hodnoty a konečně znak **US** má nulové hodnoty, když povodeň 18 má nízkou hodnotu.

4.shluk obsahuje povodně 15, 6, 13, 12, 14, 1. Znak **Qkul** vykazuje nadprůměrné až vysoké hodnoty, **Qo** podprůměrné hodnoty, znaky **Psr** a **UPS** vykazují velmi nízké až podprůměrné hodnoty, znak **MAXsr** velmi nízké až podprůměrné hodnoty, vyjma povodně 13 s nadprůměrnou hodnotou, a znak **US** má nadprůměrné až velmi vysoké hodnoty.

Samostatným objektem, který zcela vybočuje jako odlehlá hodnota, je povodeň 9, což je letní povodeň s extrémními hodnotami **Qkul, INTp** a **MAXsr**, dále s velmi vysokými hodnotami **Psr, UPS** a **INTvz** a naopak s podprůměrnou hodnotou **Qo**.

7. Závěr

Původních 13 sledovaných znaků lze zredukovat na tři latentní proměnné, zvané faktory nebo hlavní komponenty. Do prvního nejvýznamnějšího faktoru FA1 se promítají srážkové ukazatele a tání sněhu, druhý faktor FA2 souvisí se strmostí povodňové vlny a průtokem před povodní. Vysvětlený rozptyl v datech v prostoru tří nových faktorů je necelých 60 %, to znamená, že více než 40 % variability v datech byly vyhodnoceny jako šum. Za důvod vysoké hodnoty šumu, lze považovat jednak charakter dat, kdy povodňové stavy jsou sami o sobě extrémními případy, a jednak malý počet sledovaných povodních ku počtu původních ukazatelů. Optimum bývá okolo 20 objektů na 1 sledovaný znak a minimální poměr by měl být proto 5 ku 1. Z hlediska analýzy objektů se povodně zřetelně dělí na dva významné shluky: povodně v zimním a letním období. Tyto významně odlišné skupiny by bylo optimální vyhodnocovat odděleně. V tom případě by ovšem poměr objektů a znaků byl ještě nižší.

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

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Kendall M. G., Stuart A.: *The Advanced Theory of Statistics*, Vol. III. New York 1966.
[3] James W., Stein C.: *Estimation with Quadratic Loss*, Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.
[4] Guanadeskian R., Kettenring J. R.: *Biometrics* **28**, 80 (1972).
[5] Campbell N. A.: *Appl. Statist.*, 29, 231 (1980).
[6] Hu J., Skrabal P., Zollinger H.: *Dyes and Pigments*, **8**, 189 (1987).
[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.
[8] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.
[9] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[10] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.
[11] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[12] Andrews D. F.: *Biometrics*, **28**, 125 (1972).
[13] Kulkarni S. R., Paranjape S. R.: *Commun. Statist.*, **13**, 2511 (1984).
[14] Guanadeskian R.: *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York 1977.
[15] Kleiner B., Hartigan J. A., *J. Amer. Statist. Assoc.*, **76**, 260 (1981).
[16] Kres H.: *Statistical Tables for Multivariate Analysis*. Springer, New York 1983.
[17] Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.
[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: *Z. Anal. Chem.*, **327**, 679 (1987).
[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: *Technometrics*, **24**, 139 (1982).
[20] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982
[21] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy*, SNTL Praha 1981
[22] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester 1992.
[23] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[24] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).
[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004.
[27] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.
[28] Thomas E. V., *Anal. Chem.*, **66** (1994) 795A-804A.
[29] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.
[30] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat*, Univerzita Pardubice, 1996.
[31] Meloun M. , Militký J., *Kompendium statistického zpracování dat*, Academia Praha 2002, Academia Praha 2006.

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

Ing. Jindřich Freisleben
ČHMÚ
Na Šabatce 17
143 06 Praha 4 – Komořany
email: freisleben@chmi.cz, telefon: 244 032 331

Multivariate Statistical Analysis of Floods on the River Sázava within 1961 – 2000 (Meloun, M., Freisleben, J.)

Key Words
PCA - FA - CA - CM - Cluster Analysis - Dendrogram - Cattel´s index diagram -

Multivariate statistical analysis is based on the latent variables which are formed as the linear combination of original variables. The source data matrix contains here objects in 25 rows (floods) and variables (properties of floods) in 13 columns. Before data treatment the data are scaled. Similarity of objects and variables is considered on base on Mahalanobis distance in the 13-dimensional space. The principal components analysis PCA reduces dimensionality and presents floods in two or three dimensions. The plot of components weight shows hidden structure among variables while the scatterplot shows the hidden structure of objects. The cluster analysis leads to clusters which may be plotted in dendrogram. There are two dendrograms available, the dendrogram of variables properties) and the dendrogram of objects (floods).