Shluková analýza a její aplikace v hydrologii

Cíl shlukové analýzy

Cílem shlukové analýzy je roztřídění n objektů, z nichž každý je popsán p znaky, do několika pokud možno stejnorodých (homogenních) skupin (shluků, clusterů). Požadujeme, aby objekty uvnitř shluků si byly podobné co nejvíce, zatímco objekty z různých shluků co nejméně. Přesný počet shluků většinou není známý.

Shluková analýza nachází uplatnění v celé řadě oborů, např. v biologii. U n populací změříme p biometrických charakteristik a zjišťujeme, zda určité skupiny populací tvoří shluky.

Shluková analýza je ovšem průzkumovou metodou a měla by sloužit jako určité vodítko při dalším zpracování dat.

Podobnost objektů

Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost. Nechť k-tý objekt je popsán vektorem pozorování $\mathbf{x}_k = (x_{k1}, ..., x_{kp})^T$ a l-tý objekt vektorem $\mathbf{x}_l = (x_{l1}, ..., x_{lp})^T$. Euklidovská vzdálenost k-tého a l-tého objektu:

$$d_{kl} = \sqrt{\sum_{j=1}^{p} \left(x_{kj} - x_{lj}\right)^2} .$$

Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

Hierarchické shlukování

Při aplikacích shlukové analýzy se nejčastěji používá aglomerativní hierarchická procedura. Její princip spočívá v postupném slučování objektů, a to nejprve nejbližších a v dalších krocích pak stále vzdálenějších.

Algoritmus:

1. krok: Každý objekt považujeme za samostatný shluk.

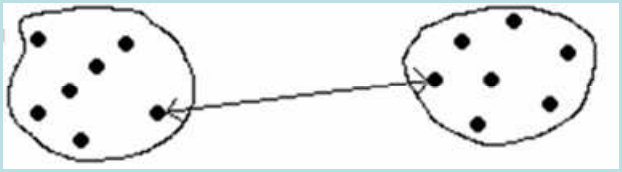
2. krok: Najdeme dva shluky, jejichž vzdálenost je minimální.

3. krok: Tyto dva shluky spojíme v nový, větší shluk a přepočítáme matici vzdáleností. Její řád se sníží o 1. Vrátíme se na 2. krok.

Funkce algoritmu končí, až jsou všechny objekty spojeny do jediného shluku.

Vzdálenost mezi shluky se počítá různými způsoby. Uvedeme tři z nich.

a) Metoda nejbližšího souseda: Vzdálenost mezi dvěma shluky je minimem ze všech vzdáleností mezi jejich objekty.



b) Metoda nejvzdálenějšího souseda: Vzdálenost mezi dvěma shluky je maximem ze všech vzdáleností mezi jejich objekty.

c) Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.

Výhody a nevýhody uvedených tří metod

Metoda průměrné vazby

Nevýhoda: řetězový efekt (spojují se shluky, jejichž dva objekty jsou sice nejbližší, ale vzhledem k většině ostatních objektů nejde o nejbližší shluky)

Výhody: Je invariantní k monotónním transformacím matice podobností a není ovlivněna vazbami v datech. První vlastnost, invariantnost k monotónní transformaci, je celkem důležitá, neboť téměř všechny další hierarchické aglomerativní metody tuto vlastnost nemají. To znamená, že metoda nejbližšího souseda je jedna z mála metod, které nejsou ovlivněny žádnou transformací dat.

Metoda nejvzdálenějšího souseda

Výhoda: odpadá řetězový efekt, vede k tvorbě relativně malého počtu poměrně kompaktních shluků.

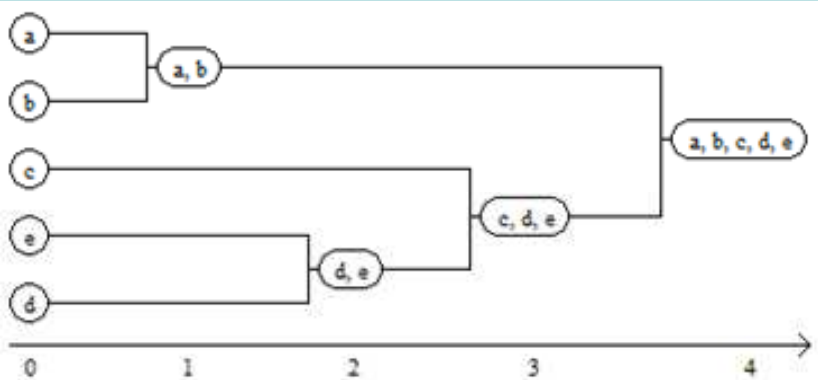
Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.

Vede k podobným výsledkům jako metoda nejvzdálenějšího souseda.

Tyto tři metody nevyžadují původní data, stačí jim matice vzdáleností.

Výsledky aglomerativní hierarchické procedury se zpravidla znázorňují pomocí dendrogramu. Je to graficky znázorněná posloupnost dvojic $\{(\nu_1, S^{(1)}), \ldots, (\nu_n, S^{(n)})\}$, kde $\{\nu_i\}_{i=1}^{n}$ je neklesající posloupnost úrovní spojování a $S^{(i)}$ je roztřídění objektů odpovídající úrovni ν_i, i = 1, ..., n.

Příklad dendrogramu:



V levém sloupci jsou jednotlivé objekty, další sloupce reprezentují shluky, do nichž byly objekty zařazeny a délky čar představují vzdálenosti mezi shluky.

Poznámka:

Hierarchická shluková analýza může být použita nejen na shlukování objektů, ale též na shlukování znaků.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, z něhož je zjevná struktura objektů ve shlucích.

Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (všeobecně m-tice) znaků, které si jsou velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku, si jsou značně podobné a jsou tudíž vzájemně nahraditelné. To má značný význam při plánování experimentu - některé vlastnosti či znaky není zapotřebí vůbec zjišťovat či měřit, protože jsou snadno nahraditelné jinými znaky a nemají velkou vypovídací hodnotu

Kofenetický koeficient korelace

Různé shlukovací procedury mohou poskytovat různé výsledky. K posouzení shody mezi maticí vzdáleností objektů a výsledkem dané shlukovací metody je možno použít např. kofenetický koeficient korelace. Posuzuje míru shody mezi maticí vzdáleností objektů a výsledkem dané shlukovací metody. Je to koeficient korelace mezi n(n-1)/2 prvky umístěnými nad (nebo pod) hlavní diagonálou matice vzdáleností a odpovídajícími prvky kofenetické matice. Přitom (i,j)-tý prvek této matice je definován jako ta vzdálenost i-tého a j-tého objektu, při níž jsou tyto objekty poprvé spojeny do jednoho shluku. Této vzdálenosti se říká kofenetická vzdálenost.

Z uvažovaných shlukovacích metod pak vybereme tu, která poskytuje nejvyšší kofenetický koeficient korelace.

Upozornění: Systém STATISTICA bohužel neposkytuje kofenetický koeficient korelace. Je možno ho získat pomocí systému MATLAB.

Postup při provádění shlukové analýzy ve STATISTICE a MATLABu

Načteme datový soubor.

Proměnné, které vstoupí do shlukování, znázorníme pomocí krabicových diagramů. Pokud se nám jeví, že variabilita jednotlivých proměnných je příliš odlišná, provedeme standardizaci.

Na data (či standardizovaná data) aplikujeme postupně metodu nejbližšího souseda, nejvzdálenějšího souseda metodu průměrné vazby.

Vybereme tu metodu, která poskytne nejvyšší kofenetický koeficient korelace. Ten spočítáme v MATLABu.

Do matice X uložíme zkoumaný datový soubor.

Y = pdist(X, ’euclid’) … poskytne řádkový vektor obsahující prvky nad hlavní diagonálou matice euklidovských vzdáleností.

Z = linkage(Y,’single’) … poskytne matici o n-1 řádcích a 3 sloupcích, která obsahuje informace potřebné pro sestrojení dendrogramu (parametr single je pro metodu nejbližšího souseda, pro metodu nejvzdálenějšího souseda je complete, pro metodu průměrné vazby average a pro Wardovu metodu ward).

c = cophenet(Z,Y) … poskytne kofenetický koeficient korelace.

dendrogram(Z) … vykreslí se dendrogram pro výsledky zvolené hierarchické aglomerativní procedury.

Vhodný počet shluků stanovíme podle průměrné hodnoty siluetové funkce. Nabývá hodnot od -1 do 1 a interpretuje se takto:

Hodnoty s(i)	interpretace
0,71 až 1	i-tý objekt se silně váže k danému shluku
0,51 až 0,7	i-tý objekt je dobře zařazen do shluku
0,26 až 0,5	i-tý objekt se slabě váže k danému shluku
-1 až 0,25	i-tý objekt je pravděpodobně chybně zařazen

Získání hodnot siluetové funkce v MATLABu:

Do sloupcového vektoru ID uložíme čísla shluků, do nichž jsou jednotlivé objekty zařazeny. Tento vektor získáme ve STATISTICE v dialogu pro hierarchickou shlukovou analýzu, kde na záložce Detaily zvolíme Uložit klasifikaci a provedeme řez dendrogramem na zvolené úrovni spojování.

[S,H]=silhouette(X,ID) … poskytne hodnoty siluetové funkce pro všechny zkoumané objekty, jejichž roztřídění do shluků je dáno vektorem ID a nakreslí siluetový graf.

m=mean(S) … průměrná hodnota siluetové funkce.

Vybereme takový počet shluků, pro nějž je průměrná hodnota siluetové funkce největší.

Roztřídění, které poskytne hierarchická shluková analýza, ověříme pomocí metody k-průměrů.

Metoda k-průměrů

Chceme-li verifikovat výsledek dané hierarchické shlukovací metody, můžeme tak učinit např. pomocí metody k-průměrů, což je nehierarchická shlukovací procedura, která vychází z následujícího algoritmu:

Algoritmus metody k-průměrů

1. krok: Stanovíme počáteční rozklad množiny n objektů do k shluků. Rozklad zpravidla volíme náhodně.

2. krok: Určíme výběrové centroidy v aktuálních shlucích. (Výběrovým centroidem shluku rozumíme hypotetický objekt, jehož vektor pozorování je roven vektoru výběrových průměrů všech objektů patřících do tohoto shluku.)

3. krok: Pro všechny objekty spočteme jejich vzdálenosti od všech výběrových centroidů. Objekt zařadíme do toho shluku, k jehož výběrovému centroidu má nejblíže. Pokud nedošlo v tomto kroku k žádnému přesunu, považujeme aktuální shluky za definitivní, jinak se vracíme ke 2. kroku.

Vliv, který mají jednotlivé proměnné na zařazení objektů do shluků, posoudíme pomocí tabulky ANOVA.

Příklad

Pro analýzu chování průtoků jsme vybrali stanici Kychová, která se nachází v experimentálním povodí říčky Kychovky v Javorníkách. Stanice leží v zalesněné oblasti, která není příliš ovlivněna antropogenní činností. Průtoky na Kychovce jsou měřeny od roku 1930 až do současnosti. Zde se omezíme na zpracování dat z let 2000 – 2014.

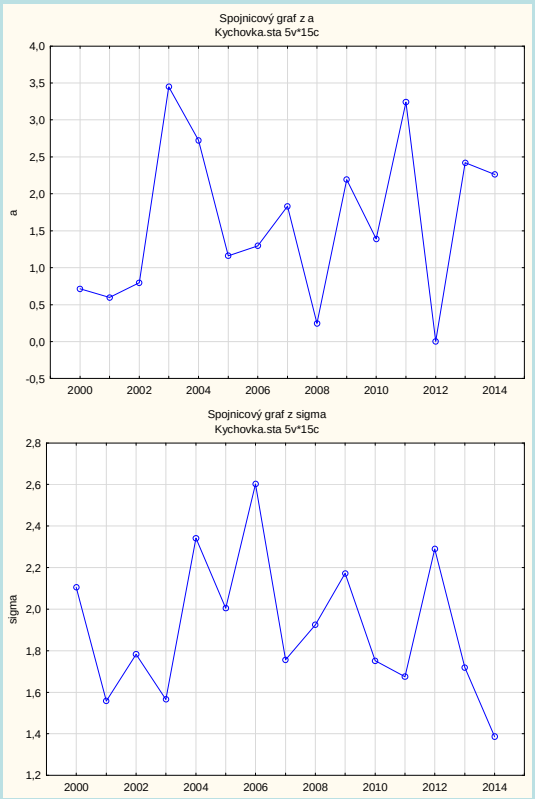
Prvotní analýzy spočívaly v odhadu parametrů rozložení LN5 všech 365 průměrných denních průtoků pro každý hydrologický rok (hydrologický rok trvá od 1.11. do 31.10.) zvlášť. Aby bylo možno odhadnuté parametry a, b, μ, σ, y_0 porovnávat, byly průtoky pro každý jednotlivý rok normovány na průměr 1.

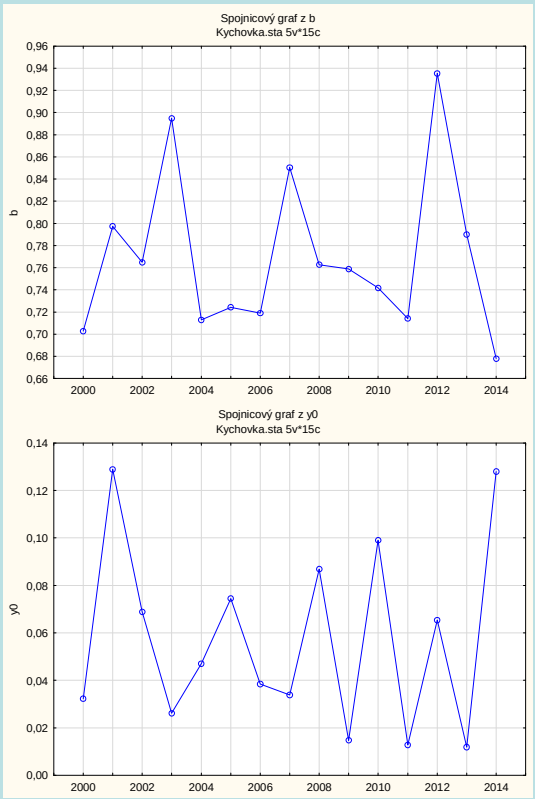
Pomocí odhadnutých parametrů proveďte shlukovou analýzu pro jednotlivé roky.

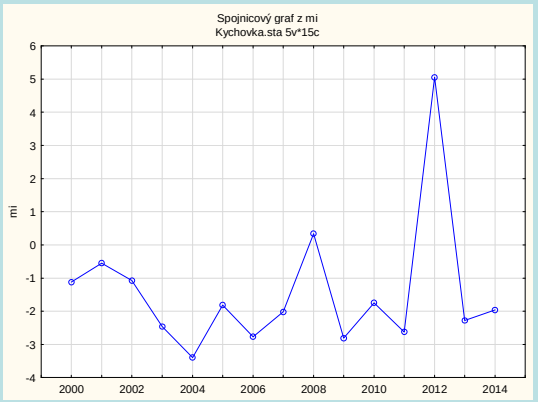
Řešení:

Zobrazíme časový průběh pěti parametrů a, b, μ, σ, y_0.

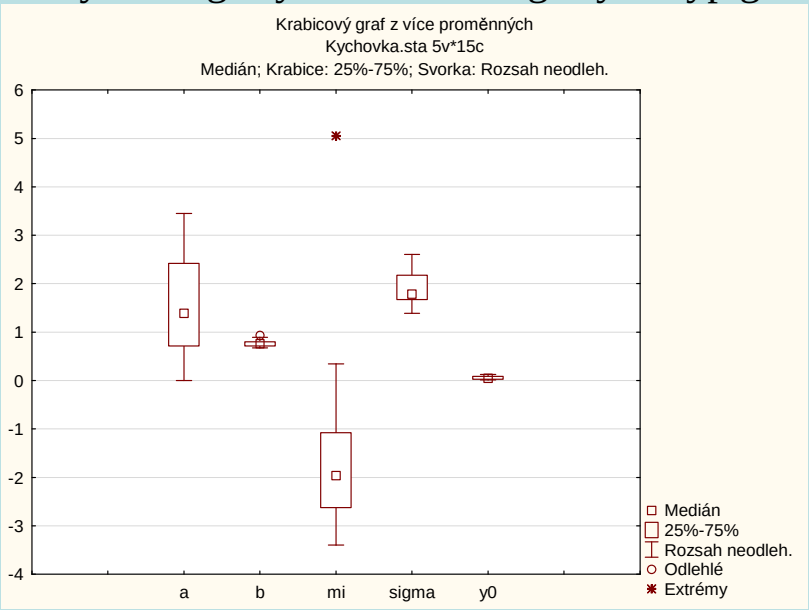
Grafy – 2D grafy – Spojnicové grafy (proměnné) – Proměnné – Vše – OK.





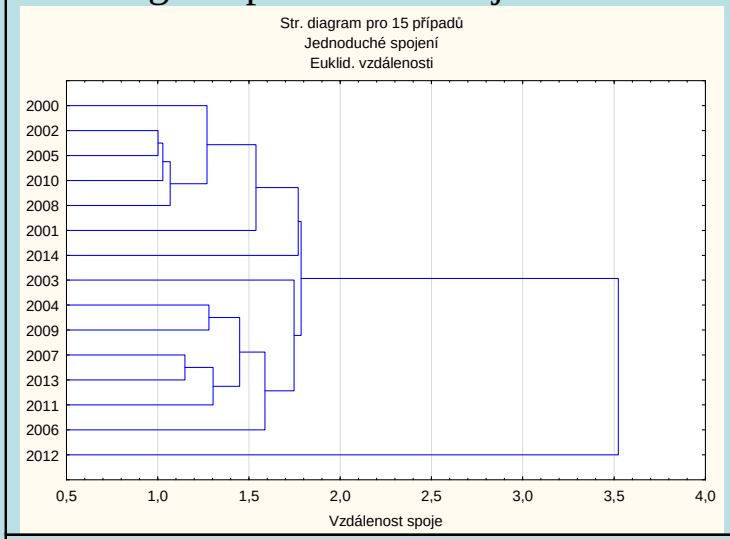


Dále vytvoříme krabicové diagramy:
Grafy – 2D grafy – Krabicové grafy – Typ grafu Vícenásobný – Proměnné – Vše – OK - OK.

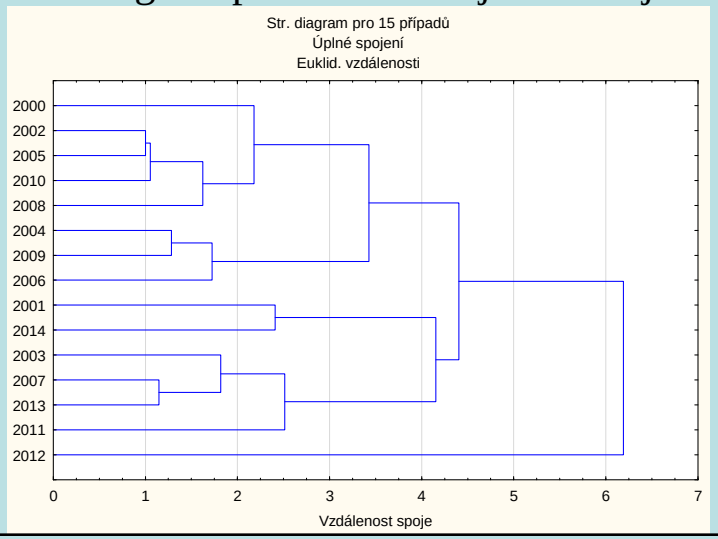


Vzhledem k velmi rozdílné variabilitě proměnných budeme pracovat se standardizovanými daty.
Data – Standardizovat – Proměnné – Vše, Případy – ALL – OK.
Pro standardizované proměnné provedeme shlukovou analýzu s euklidovskou vzdáleností a třemi metodami: nejbližšího souseda, nejvzdálenějšího souseda a průměrné vazby. Vytvoříme dendrogramy.
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza Spojování (hierarchické shlukování) – OK – Proměnné – Vše, OK, Detaily - Shlukovat případy (řádky) – Pravidlo slučování: Jednoduché spojení – Míry vzdálenosti: Euclidovské vzdálenosti – OK – Horizontální graf hierarch. stromu. Euklidovská vzdálenost a metoda nejbližšího souseda je nastavena implicitně. Pro další dvě metody změníme Pravidlo slučování z Jednoduchého spojení na Úplné spojení resp. Nevážený průměr skupin dvojic.

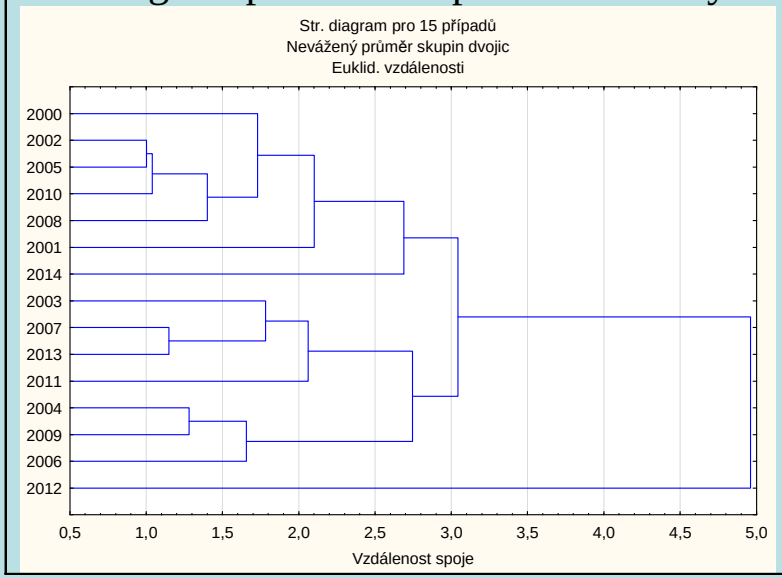
Dendrogram pro metodu nejbližšího souseda



Dendrogram pro metodu nejvzdálenějšího souseda



Dendrogram pro metodu průměrné vazby



Uvedené metody dávají poněkud odlišné výsledky. Shodu mezi maticí vzdáleností a dendrogramem posoudíme pomocí kofenetických koeficientů korelace. Tyto koeficienty byly vypočítány pomocí systému MATLAB.
Do matice X uložíme soubor standardizovaných dat (pozor – v MATLABu se používají desetinné tečky).
Y = pdist(X, ’euclid’) … poskytne řádkový vektor obsahující prvky nad hlavní diagonálou matice euklidovských vzdáleností.
Z = linkage(Y,’single’) … poskytne matici o n-1 řádcích a 3 sloupcích, která obsahuje informace potřebné pro sestrojení dendrogramu (parametr single je pro metodu nejbližšího souseda, pro metodu nejvzdálenějšího souseda je complete, pro metodu průměrné vazby average a pro Wardovu metodu ward).
c = cophenet(Z,Y) … spočítá kofenetický koeficient korelace.

metoda	kofenetický koeficient
nejbližšího souseda	0,7820
nejvzdálenějšího souseda	0,7808
průměrné vazby	0,8245

Nejvyšší kofenetický koeficient poskytla metoda průměrné vazby, tedy nadále budeme uvažovat její výsledky.

Ze vzhledu dendrogramu vyplývá, že roky 2000 – 2014 utvoří buď 3 nebo 4 shluky - větší počet shluků už by poskytl příliš roztříštěnou informaci. Rozhodnutí o počtu shluků založíme na průměrné šířce rozkladu, tj. na průměrné hodnotě siluetové funkce při roztřídění objektů do určitého počtu shluků.

Nejprve musíme získat čísla shluků, do nichž jsou jednotlivé roky zařazeny. Postup ukážeme ve STATISTICE pro 3 shluky.
V dialogu pro hierarchickou shlukovou analýzu, kde na záložce Detaily zvolíme Uložit klasifikaci a provedeme řez dendrogramem na zvolené úrovni spojování, abychom dostali tři shluky.

Sloupec se zařazením uložíme do MATLABu jako proměnnou ID.
Zavoláme funkci silhouette:
[S,H]=silhouette(X,ID) … poskytne hodnoty siluetové funkce pro všechny zkoumané objekty, jejichž roztřídění do shluků je dáno vektorem ID a nakreslí siluetový graf.
Spočteme průměrnou hodnotu siluetové funkce:
m=mean(S) … průměrná hodnota siluetové funkce.

Výsledky pro 3 shluky:		Siluetová funkce	Výsledky pro 4 shluky:		Siluetová funkce
	Zařazení do klastrů (Kychovka.sta) Spojovací vzdálenost = 3,00000 Nevážený průměr skupin dvojic Euklid. vzdálenosti			Zařazení do klastrů (Kychovka.sta) Spojovací vzdálenost = 2,71939 Nevážený průměr skupin dvojic Euklid. vzdálenosti	
	Zařazení do klastrů			Zařazení do klastrů	
2000	2	0,1452	2000	4	-0,3252
2001	2	0,6691	2001	4	0,6394
2002	2	0,6246	2002	4	0,6203
2003	3	0,4117	2003	3	0,6816
2004	3	0,3883	2004	2	0,7300
2005	2	0,5185	2005	4	0,2760
2006	3	0,0840	2006	2	0,7269
2007	3	0,2415	2007	3	0,5145
2008	2	0,6737	2008	4	0,6355
2009	3	0,5747	2009	2	0,4666
2010	2	0,6796	2010	4	0,6765
2011	3	0,4745	2011	3	0,3120
2012	1	1,0000	2012	1	1,0000
2013	3	0,5215	2013	3	0,6316
2014	2	0,4701	2014	4	0,4206
Průměrná hodnota siluetové funkce m = 0,4985			Průměrná hodnota siluetové funkce m = 0,5338		

Průměr hodnot siluetové funkce je vyšší při zařazení roků do 4 shluků:
2012
2003, 2007, 2011, 2013
2004, 2006, 2009
2000, 20001, 2002, 2005, 2008, 2010, 2014

Provedení shlukové analýzy metodou k-průměrů

Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Shlukování metodou k-průměrů – OK – Proměnné Vybrat vše – OK – Shlukovat: Případy (řádky), na záložce Detaily zadáme počet shluků 4 – OK. Na záložce Detaily vybereme Členy shluků a vzdálenosti. Dostaneme 4 tabulky, které obsahují roky v 1. až 4. shluku a vzdálenosti roků od středu shluku:

	Členy shluku číslo 1 (Kychovka.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 1 příp.
	Vzdálen.
2012	0,00

	Členy shluku číslo 2 (Kychovka.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 4 příp.
	Vzdálen.
2003	0,605847
2007	0,482580
2011	0,643978
2013	0,223465

	Členy shluku číslo 3 (Kychovka.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 4 příp.
	Vzdálen.
2000	0,601381
2004	0,489514
2006	0,441182
2009	0,402500

	Členy shluku číslo 4 (Kychovka.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 6 příp.
	Vzdálen.
2001	0,582322
2002	0,374076
2005	0,486456
2008	0,561880
2010	0,190606
2014	0,880588

Ověření výsledků hierarchické shlukovací procedury metodou k-průměrů pro k = 4:

(2012), (2003, 2007, 2011, 2013), (2000, 2004, 2006, 2009), (2001, 2002, 2005, 2008, 2010, 2014).

Na rozdíl od metody průměrné vazby byl rok 2000 zařazen do shluku č. 3. Je to správně?

Znovu spočítáme průměr hodnot siluetové funkce, dostaneme 0,5559, to je víc než 0,5338.

Výsledné roztřídění:

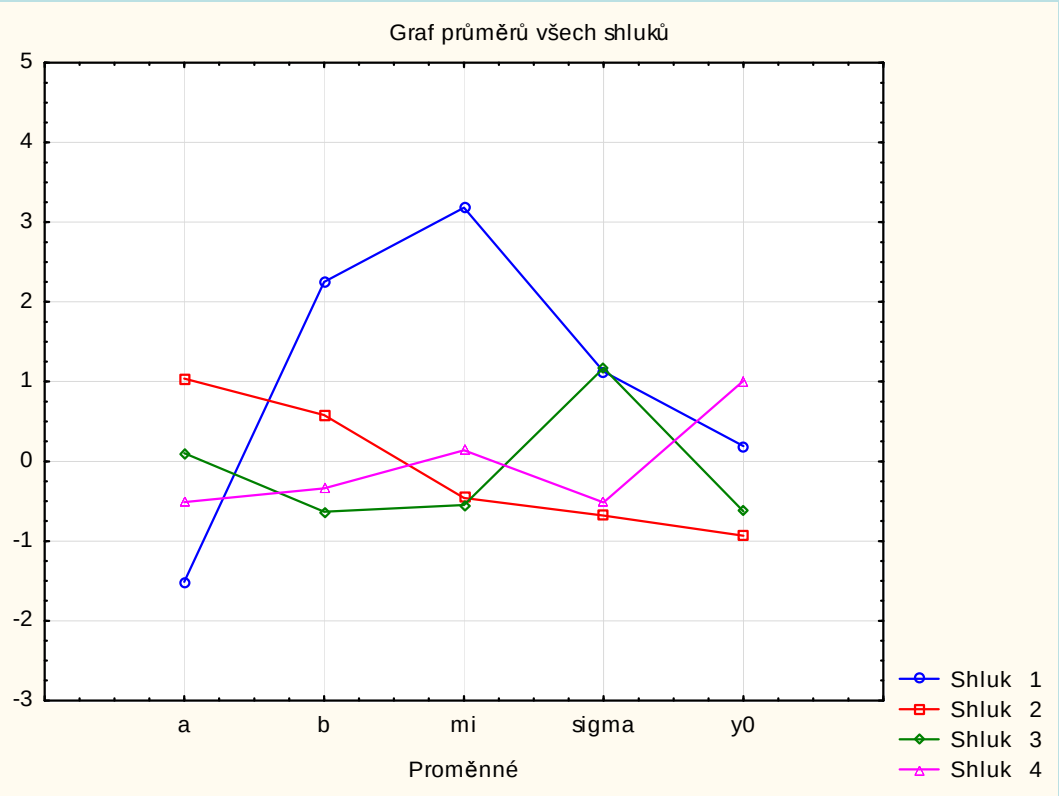
Shluk č. 1 … rok 2012

Shluk č. 2 … roky 2003, 2007, 2011, 2013

Shluk č. 3 … roky 2000, 2004, 2006, 2009

Shluk č. 4 … roky 2001, 2002, 2005, 2008, 2010, 2014

Průměry odhadnutých parametrů v jednotlivých shlucích:



Shluk č. 1 … rok 2012
Shluk č. 2 … roky 2003, 2007, 2011, 2013
Shluk č. 3 … roky 2000, 2004, 2006, 2009
Shluk č. 4 … roky 2001, 2002, 2005, 2008, 2010, 2014

Posouzení vlivu jednotlivých odhadnutých parametrů na zařazení roků do shluků pomocí analýzy rozptylu:

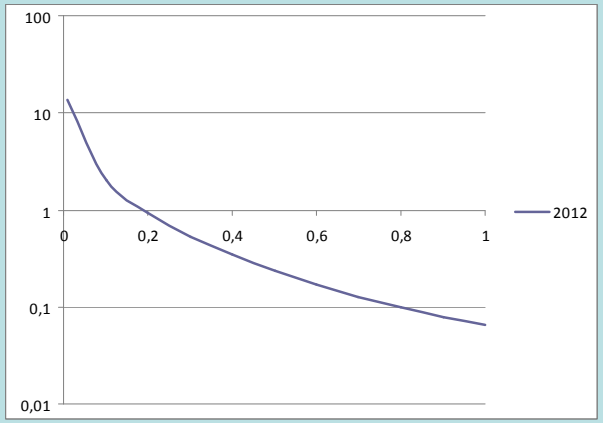
parametr	skupinový součet čtverců	stupně volnosti	reziduální součet čtverců	stupně volnosti	F	p-hodnota
a	8,23124	3	5,768759	11	5,23184	0,017352
b	8,70492	3	5,295082	11	6,02786	0,011069
μ	12,27997	3	1,720027	11	26,17782	0,000026
σ	10,22023	3	3,779765	11	9,91442	0,001847
y_0	11,23535	3	2,764647	11	14,90110	0,000344

Všechny odhadnuté parametry jsou významné na hladině významnosti 0,05, největší vliv má μ , podstatně menší y_0 a nejmenší vliv mají σ, b, a.

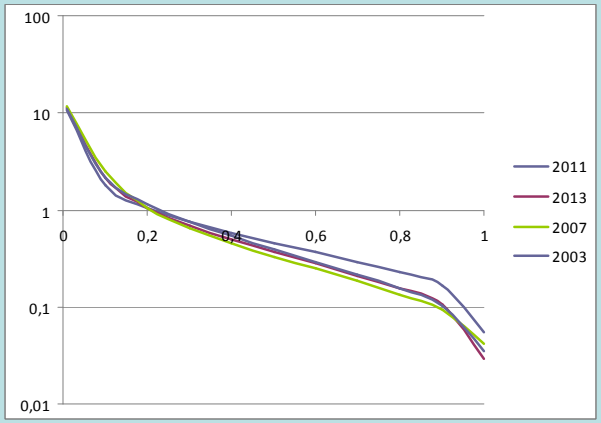
Hydrologická interpretace výsledků shlukové analýzy

Pro každý shluk byly na základě odhadnutých parametrů a, b, μ, σ, y_0 sestrojeny křivky překročení pro roky zařazené do daného shluku.

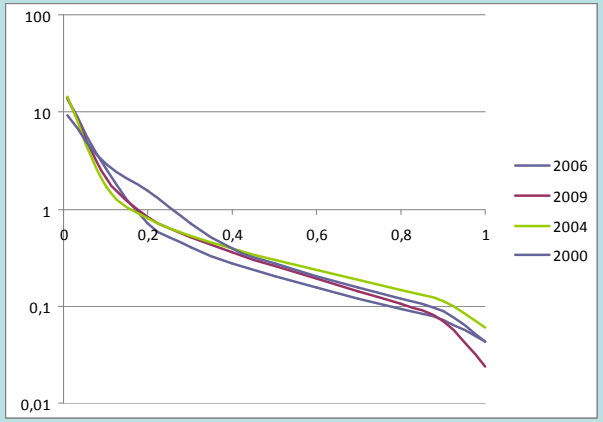
Shluk č. 1



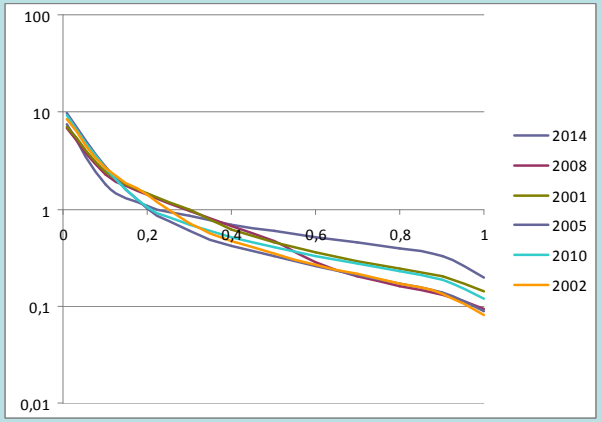
Shluk č. 2



Shluk č. 3



Shluk č. 4



Charakteristiky shluků:
Shluk č. 1 – obsahuje rok, v němž se v létě nevyskytla suchá období
Shluk č. 2 – zahrnuje roky s dlouhými suchými obdobími
Shluk č. 3 – sem patří středně suché roky
Shluk č. 4 – sdružuje středně vlhké roky