

# Statistické metody pro detekci změny ročního chodu průtokových řad

Hana Horáková<sup>1</sup>, Daniela Jarušková<sup>2</sup>, Ladislav Satrapa<sup>3</sup>

## 1 Úvod

Průtokové řady, které obecně vykazují velmi vysokou variabilitu, se obvykle považují za stacionární. Přesto se zdá, že i v nich může docházet ke změnám, které se však spíše projevují v chování denních řad než v řadách vzniklých průměrováním za delší časové období, např. za rok či měsíc. Typickým jevem je např. častější výskyt velmi vysokých denních hodnot či časový posun jarních kulminací způsobených táním sněhu. Příspěvek se soustřeďuje na statistické zkoumání změn ročního chodu. Teoreticky by bylo možno zkoumat přímo chování řady denních průměrů a použít statistické testy, které zkoumají, zda existuje změna ve střední hodnotě vektoru, který má 365 složek, kde každá složka reprezentuje jeden kalendářní den. Vzhledem k tomu, že měření byla prováděna 50-90 let, nelze spolehlivě takovou změnu odhalit. Je tedy třeba snížit dimenzi problému tím, že vektor denních průtoků nahradíme vektorem s nižší dimenzí. K tomu jsme použili rozvoj do Fourierovy řady a metodu hlavních komponent.

## 2 Dvouvýběrový problém

Ve dvouvýběrovém problému jsou studovány dvě skupiny dat  $x_1, \dots, x_n$  a  $y_1, \dots, y_m$ , kde  $\{x_i\}$  a  $\{y_j\}$  jsou  $p$  – dimenzionální vektory. Předpokládejme, že  $x_1, \dots, x_n$  jsou realizace náhodného výběru  $X_1, \dots, X_n$  z rozdělení daného distribuční funkcí  $F_1(x)$  a  $y_1, \dots, y_m$  jsou realizace náhodného výběru  $Y_1, \dots, Y_m$  z rozdělení daného distribuční funkcí  $F_2(x)$ . Dvouvýběrový problém spočívá v rozhodnutí, zda  $F_1(x)=F_2(x)$ , což znamená, že rozdělení, z nichž pocházejí  $\{x_i\}$  a  $\{y_j\}$ , jsou stejná. Více informací o dvouvýběrovém problému se lze dozvědět v literatuře [1]. Tuto úlohu můžeme v rámci matematické statistiky studovat neparametricky nebo parametricky. V případě parametrického přístupu předpokládáme, že  $F_1(x)$  i  $F_2(x)$  patří ke stejné parametrické třídě a liší se pouze v jednom nebo více parametrech. Nejjednodušší situace nastává, když předpokládáme, že obě rozdělení jsou  $p$ -dimenzionální normální rozdělení se shodnou varianční maticí  $\Sigma$ . V tom případě nám jde o rozhodnutí, zda se shodují vektory středních hodnot  $\mu_x$  a  $\mu_y$ . V rámci teorie testování statistických hypotéz se jedná o testování statistické hypotézy  $H_0: \mu_x = \mu_y$  proti alternativě  $A: \mu_x \neq \mu_y$ . Nejlepším nestranným odhadem vektoru  $\mu_x$  je  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)'$ , kde  $\bar{x}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n x_{ij}$ ,  $j=1, \dots, p$ , a nejlepším nestranným odhadem vektoru  $\mu_y$  je  $\bar{y} = (\bar{y}_1, \dots, \bar{y}_p)'$ , kde  $\bar{y}_{\cdot j} = \frac{1}{m} \sum_{i=1}^m y_{ij}$ ,  $j=1, \dots, p$ . Pokud bychom znali varianční matici  $\Sigma$ , pak testová statistika, která se používá pro testování  $H_0$  proti  $A$ , má tvar

$$\left(\frac{nm}{n+m}\right) (\bar{x} - \bar{y})' \Sigma^{-1} (\bar{x} - \bar{y}) = \left(\frac{nm}{n+m}\right) \sum_{j=1}^p \sum_{j'=1}^p (\bar{x}_{\cdot j} - \bar{y}_{\cdot j}) v_{jj'} (\bar{x}_{\cdot j'} - \bar{y}_{\cdot j'}),$$

<sup>1</sup> Hana Horáková, Ing. Mgr., ČVUT v Praze, Fakulta stavební, Katedra matematiky, Thákurova 7, Praha 6, 166 29, horakovah@mat.fsv.cvut.cz

<sup>2</sup> Daniela Jarušková, Prof., RNDr., CSc., ČVUT v Praze, Fakulta stavební, Katedra matematiky, Thákurova 7, Praha 6, 166 29, jarus@mat.fsv.cvut.cz

<sup>3</sup> Ladislav Satrapa, Doc., Ing., CSc., ČVUT v Praze, Fakulta stavební, Katedra hydrotechniky, Thákurova 7, Praha 6, 166 29, satrapa@fsv.cvut.cz

$\Sigma^{-1} = \|v_{jj'}\|_{i,j=1}^p$ , kde  $v_{jj'}$  je prvek matice  $\Sigma^{-1}$ . Tato testová statistika má za platnosti  $H_0$  rozdělení  $\chi^2$  o  $p$  stupních volnosti. Ve většině případů však matici  $\Sigma$  neznáme a musíme ji nahradit odhadem. Odhad můžeme provést buď za platnosti  $H_0$  nebo za platnosti  $A$ . Pokud platí  $H_0$ , můžeme společný vektor středních hodnot odhadnout pomocí  $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_p)'$ , kde  $\bar{\mu}_j = \frac{1}{n+m} (\sum_{i=1}^n x_{ij} + \sum_{i=1}^m y_{ij})$ . Odhad matice  $\Sigma$  za platnosti  $H_0$ , tj. matice  $\hat{\Sigma}_{H_0}$ , má prvky  $\hat{\sigma}_{jj'}(H_0) = \frac{\sum_{i=1}^n (x_{ij} - \bar{\mu}_j)(x_{ij'} - \bar{\mu}_{j'}) + \sum_{i=1}^m (y_{ij} - \bar{\mu}_j)(y_{ij'} - \bar{\mu}_{j'})}{n+m}$ ,  $j = 1, \dots, p$ ,  $j' = 1, \dots, p$ . Odhad matice  $\Sigma$  za platnosti  $A$ , tj. matice  $\hat{\Sigma}_A$ , má prvky  $\hat{\sigma}_{jj'}(A) = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_{.j})(x_{ij'} - \bar{x}_{.j'}) + \sum_{i=1}^m (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'})}{n+m}$ ,  $j = 1, \dots, p$ ,  $j' = 1, \dots, p$ . Testové statistiky  $\left(\frac{nm}{n+m}\right) (\bar{x} - \bar{y})' \Sigma_{H_0}^{-1} (\bar{x} - \bar{y})$  a  $\left(\frac{nm}{n+m}\right) (\bar{x} - \bar{y})' \Sigma_A^{-1} (\bar{x} - \bar{y})$  mají asymptoticky (pro velké  $n$  a  $m$ )  $\chi^2$  rozdělení o  $p$  stupních volnosti, viz [2].

### 3 Použití dvouvýběrového testu pro testování změny chování denního toku

Předpokládejme, že pozorujeme časovou řadu  $Z_1 = (Z_{1,1}, \dots, Z_{1,365})', \dots, Z_{2n} = (Z_{2n,1}, \dots, Z_{2n,365})'$ , kde  $2n$  je počet roků, v nichž sledujeme např. průtok řeky. Hodnota  $Z_{ij}$   $i=1, \dots, 2n$ ,  $j=1, \dots, 365$  odpovídá dennímu průtoku v  $j$ -tém dni  $i$ -tého roku sledování. Předpokládejme, že rozdělení  $Z_1, \dots, Z_{2n}$  je 365 - rovnoměrné normální rozdělení se stejnou varianční maticí. Otázkou je, zda vektory středních hodnot  $\mu_1^Z = EZ_1, \dots, \mu_{2n}^Z = EZ_{2n}$  mohou být považovány za stejné nebo zda v nich došlo ke změně. Jednou z možností, jak o daném problému rozhodnout, je použít dvouvýběrový test pro testování shody vektoru středních hodnot. Nejrozměrnější se zdá rozdělit řadu na 2 řady o stejné délce  $Z_1, \dots, Z_n$  a  $Z_{n+1}, \dots, Z_{2n}$  a testovat  $H_0: \mu_1^Z = \dots = \mu_{2n}^Z$  proti  $A: \mu_1^Z = \dots = \mu_n^Z = \mu_Z^*, \mu_{n+1}^Z = \dots = \mu_{2n}^Z = \mu_Z^{**}, \mu_Z^* \neq \mu_Z^{**}$ , kde  $\mu_i^Z = EZ_i$ . Mohli bychom teoreticky použít výše uvedenou testovou statistiku. Pokud však  $2n < 365$ , budou matice  $\hat{\Sigma}_{H_0}$ , resp.  $\hat{\Sigma}_A$  singulární, a tedy testovou statistiku není možné spočítat.

### 4 Redukce dimenzionality

Redukce dimenzionality spočívá v nahrazení vektorů  $Z_1, \dots, Z_{2n}$  vektory  $V_1, \dots, V_{2n}$ , kde  $V_i = ((c(1))'Z_i, (c(2))'Z_i, \dots, (c(k))'Z_i)$ ,  $i=1, \dots, 2n$ ,  $c(1), \dots, c(k)$  jsou vhodně zvolené vektory konstant. To znamená, že studujeme vektory, jejichž souřadnice jsou určitým způsobem zvolené lineární kombinace složek původních vektorů, přičemž se volí  $k \ll p$ .

#### 4.1 Redukce dimenzionality pomocí aproximace denního chodu Fourierovou řadou

Při této redukci volíme  $c(1)' = \left(\frac{1}{365}, \frac{1}{365}, \dots, \frac{1}{365}\right), \dots, c(2l)' = \frac{2}{365} \left(\cos \frac{2l\pi}{365}, \cos \frac{2l\pi \cdot 2}{365}, \dots, \cos 2l\pi\right)$ ,  $c(2l+1)' = \frac{2}{365} \left(\sin \frac{2l\pi}{365}, \sin \frac{2l\pi \cdot 2}{365}, \dots, \sin 2l\pi\right)$ . To znamená, že nahradíme vektor  $Z_i = (Z_{i1}, \dots, Z_{i365})'$  vektorem  $V_i = (\hat{\mu}_i, \hat{A}_{i1}, \hat{B}_{i1}, \dots, \hat{A}_{il}, \hat{B}_{il})'$ , kde  $\hat{A}_{ij} = \frac{2}{365} \sum_{t=1}^{365} \left(\cos \frac{2j\pi t}{365} \cdot Z_{it}\right)$  a  $\hat{B}_{ij} = \frac{2}{365} \sum_{t=1}^{365} \left(\sin \frac{2j\pi t}{365} \cdot Z_{it}\right)$ ,  $j = 1, \dots, l$ .

Je všeobecně známo [3], že  $\{\hat{A}_{ij}\}$ , resp.  $\{\hat{B}_{ij}\}$ , jsou odhady metodou nejmenších čtverců v modelu

$$Z_{it} = \mu_i + \sum_{j=1}^l \left( A_{ij} \cos \frac{2\pi jt}{365} + B_{ij} \sin \frac{2\pi jt}{365} \right) + e_{it}, t = 1, \dots, 365, i = 1, \dots, 2n.$$

Metodu můžeme tedy interpretovat tak, že nahradíme řadu denních hodnot v jednom roce její aproximací pomocí konečné Fourierovy řady a testujeme, zda se shodují střední hodnoty příslušných koeficientů, tj.  $\mu_1 = \dots = \mu_{2n}$ ,  $A_{11} = \dots = A_{2n,1}$ ,  $B_{11} = \dots = B_{2n,1}$ ,  $A_{1l} = \dots = A_{2n,l}$ ,  $B_{1l} = \dots = B_{2n,l}$ . Jestliže nás nezajímá změna v ročním průměru, můžeme pak testovat pouze zda  $A_{11} = \dots = A_{2n,1}$ ,  $B_{1l} = \dots = B_{2n,l}$ .

Můžeme se ptát, proč tato metoda může odhalit změnu ve střední hodnotě. Je zřejmé, že metoda bude dobře fungovat, pokud střední hodnota  $EZ_i$ ,  $i = 1, \dots, 2n$  se bude rovnat  $EZ_{it} = \mu_i + \sum_{j=1}^l \left( A_{ij} \cos \frac{2\pi jt}{365} + B_{ij} \sin \frac{2\pi jt}{365} \right)$ ,  $t = 1, \dots, 365$ , nebo alespoň se bude výrazu na pravé straně velmi blížit. Jinými slovy můžeme očekávat, že metoda bude dobře fungovat tehdy, když pro libovolné  $i = 1, \dots, 2n$  bude střední hodnota periodická hladká funkce. Záleží potom na daném problému, jak volit počet uvažovaných největších Fourierových frekvencí. V našem případě se ukázalo jako nejvhodnější volit 3 frekvence.

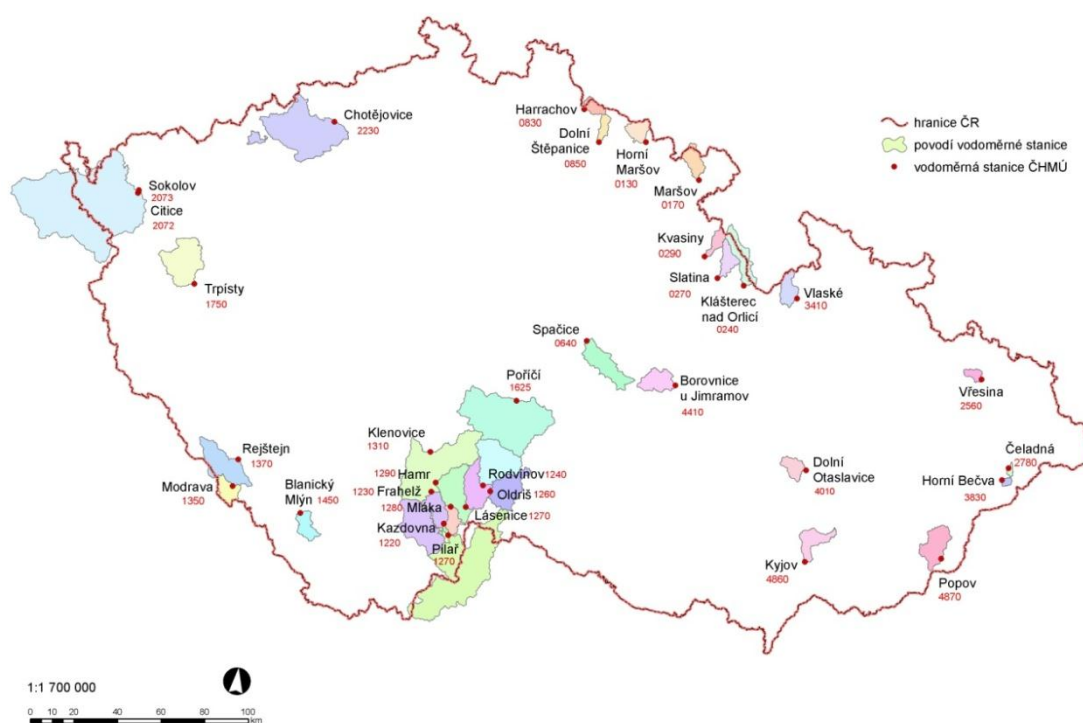
## 4.2 Redukce dimenzionality pomocí metodou hlavních komponent

Při této redukci nahradíme nejprve data  $Z_1 = (Z_{1,1}, \dots, Z_{1,365})'$ , ...,  $Z_{2n} = (Z_{2n,1}, \dots, Z_{2n,365})'$  standardizovanými hodnotami  $Z_1^S = (Z_{11} - \bar{Z}_{\cdot 1}, \dots, Z_{1,365} - \bar{Z}_{\cdot 365})'$ , ...,  $Z_{2n}^S = (Z_{2n,1} - \bar{Z}_{\cdot 1}, \dots, Z_{2n,365} - \bar{Z}_{\cdot 365})'$ , kde  $\bar{Z}_{\cdot l} = \frac{1}{2n} \sum_{i=1}^{2n} Z_{il}$ . Vektory  $c(1), \dots, c(k)$  volíme následovně:  $c(1) = \hat{u}_1, \dots, c(k) = \hat{u}_k$ , kde  $\hat{u}_1, \dots, \hat{u}_k$  jsou vlastní vektory matice  $\hat{\Sigma}_A$  odpovídající  $k$  největším vlastním číslům  $\hat{\lambda}_1 \geq \dots \geq \dots \geq \hat{\lambda}_k$ . To znamená, že nahradíme vektor  $Z_i^S = (Z_{i1}^S, \dots, Z_{i365}^S)'$  vektorem  $V_i = (\hat{u}_1' Z_i^S, \dots, \hat{u}_k' Z_i^S)'$ . Počet hlavních komponent se volí subjektivně, v našem případě jsme volili 7-12 vlastních čísel.

## 5 Vyhodnocení

Uvedené metody jsme aplikovali na 18 stanic (viz mapa na obr. 1), v nichž byly měřeny denní průtoky po dobu 50-90 let. Jedná se o následující stanice: Běla - Skuhrov (1940-2007), Blanice - Blanický Mlýn (1952-2008), Brodečka - Otaslavice (1940-2007), Čeladenka - Šance (1940-2007), Doubrava - Spačice (1952-2007), Jizera - Štěpanice (1922-2008), Kyjovka - Kyjov (1950-2008), Morava - Vlaské (1949-2008), Mumlava - Harrachov (1950-2008), Orlice - Klášterec (1938-2006), Otava - Rejštejn (1910-1920, 1930-1937, 1947-2007), Porubka - Vřesina (1952-2008), Rožnovská Bečva - Horní Bečva (1954-2008), Svatka - Borovnice (1954-2008), Úpa - Horní Maršov (1954-2008), Vlára - Popov (1955-2008), Vydra - Morava (1930-1940, 1948-2007), Zdobnice - Pěčín (1944-2007).

Na hladině významnosti  $\alpha=0,05$  vyšly při použití metody Fourierových frekvencí (pro 3 frekvence) jako významné stanice Blanice a Vydra, při volbě  $\alpha=0,10$  dále stanice Mumlava, Orlice a Otava. Při použití metody hlavních komponent jsme volili počet vlastních čísel 7-12. Při volbě hladiny významnosti  $\alpha=0,05$  lze za statisticky významné považovat hodnoty stanic Blanice (pro 7-9 vlastních čísel), Mumlava (pro 7-9 vlastních čísel) a Morava (pro 11 vlastních čísel), na hladině významnosti  $\alpha=0,10$  se výsledky významných hodnot rozšíří ještě o stanici Blanice (pro 10-12 vlastních čísel), pro Mumlavu (pro 10-12 vlastních čísel) a pro stanici Morava (pro 10 a 12 vlastních čísel).



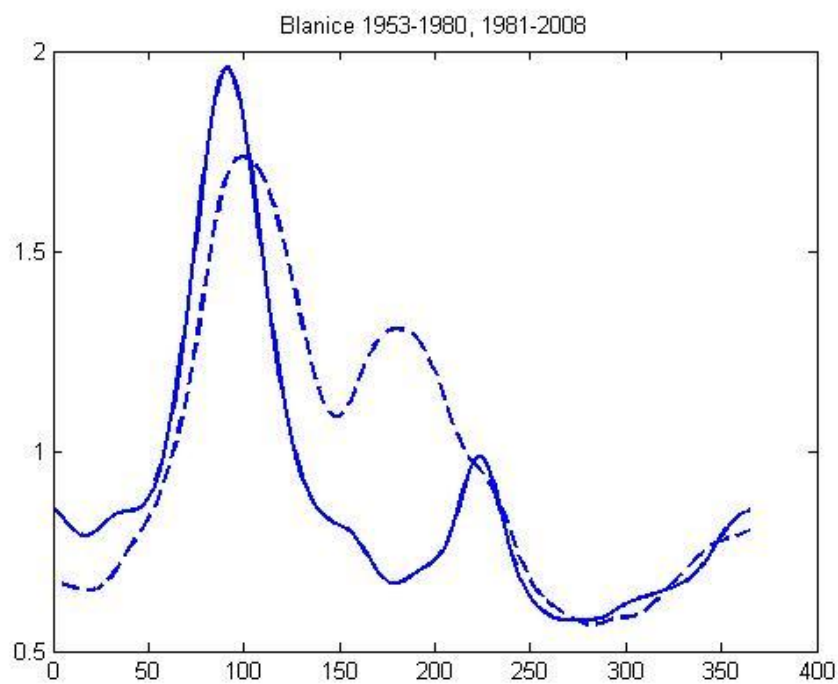
**Obr. 1** Mapa stanic

|              | <b><math>\chi^2</math> - stat. (p-hodnota)</b> |
|--------------|------------------------------------------------|
| BELA         | 5,9242 (0.4317)                                |
| BLANICE      | 19,6011 (0.0033)                               |
| BRODECKA     | 2,4446 (0.8746)                                |
| CELADENKA    | 2,9026 (0.8210)                                |
| DOUBRAVA     | 4,5397 (0.6040)                                |
| JIZERA       | 6,7926 (0.3405)                                |
| KYJOVKA      | 3,2428 (0.3405)                                |
| MORAVA       | 6,0380 (0.4189)                                |
| MUMLAVA      | 10,7054 (0.0979)                               |
| ORLICE       | 11,9143 (0.0639)                               |
| OTAVA        | 11,1898 (0.0827)                               |
| PORUBKA      | 3,0853 (0.7981)                                |
| ROZNOV.BECVA | 3,6236 (0.7275)                                |
| SVRATKA      | 4,4339 (0.6182)                                |
| UPA          | 5,4549 (0.4869)                                |
| VLARA        | 6,5175 (0.3678)                                |
| VYDRA        | 13,4804 (0.0360)                               |
| ZDOBNICE     | 5.1780 (0.5212)                                |

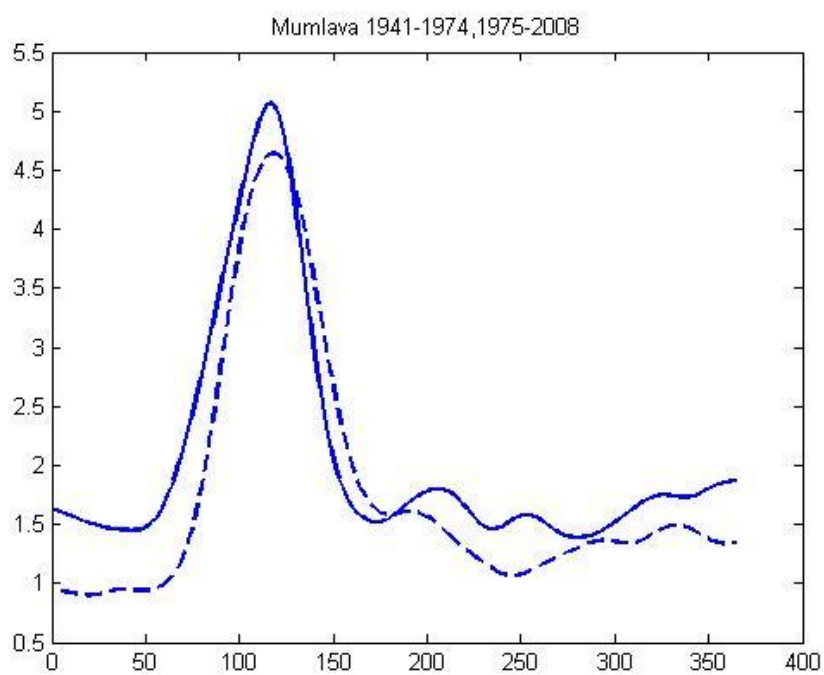
**Tab.1** Výsledné hodnoty při použití metody Fourierových frekvencí

|                           |                  |                  |                  | X <sup>2</sup> - statistika<br>(p-hodnota) |                  |                  |  |  |
|---------------------------|------------------|------------------|------------------|--------------------------------------------|------------------|------------------|--|--|
| počet vlastních<br>čísels | 7                | 8                | 9                | 10                                         | 11               | 12               |  |  |
| BELA                      | 4.1153 (0.7664)  | 4.2803 (0.8310)  | 5.9319 (0.7467)  | 8.1585 (0.6134)                            | 14.5432 (0.2044) | 14.5762 (0.2654) |  |  |
| BLANICE                   | 14.1824 (0.0480) | 17.4900 (0.0254) | 17.9354 (0.0359) | 18.2342 (0.0511)                           | 18.7648 (0.0654) | 20.4121 (0.0597) |  |  |
| BRODECKA                  | 4.2745 (0.7477)  | 4.8690 (0.7715)  | 6.0900 (0.7309)  | 6.0971 (0.8070)                            | 6.1486 (0.8633)  | 6.5323 (0.8869)  |  |  |
| CELADENKA                 | 6.0421 (0.5348)  | 7.1553 (0.5200)  | 7.9099 (0.5433)  | 8.1119 (0.6179)                            | 8.9659 (0.6250)  | 14.4888 (0.2706) |  |  |
| DOUBRAVA                  | 5.6228 (0.5844)  | 7.0848 (0.5275)  | 8.2124 (0.5129)  | 8.2159 (0.6078)                            | 8.3768 (0.6792)  | 10.1440 (0.6033) |  |  |
| JIZERA                    | 7.6297 (0.3664)  | 10.4386 (0.2356) | 10.4501 (0.3153) | 12.1304 (0.2764)                           | 12.3221 (0.3399) | 12.3369 (0.4190) |  |  |
| KYJOVKA                   | 5.0108 (0.6586)  | 5.3286 (0.7219)  | 6.8798 (0.6496)  | 8.0830 (0.6207)                            | 8.3868 (0.6783)  | 8.6413 (0.7332)  |  |  |
| MORAVA                    | 10.0564 (0.1854) | 11.7999 (0.1604) | 11.8782 (0.2203) | 16.0745 (0.0975)                           | 19.7325 (0.0491) | 20.3545 (0.0607) |  |  |
| MUMILAVA                  | 16.6833 (0.0196) | 17.1027 (0.0291) | 17.4546 (0.0421) | 17.9703 (0.0555)                           | 18.0041 (0.0815) | 18.8567 (0.0920) |  |  |
| ORLICE                    | 7.7327 (0.3567)  | 7.7371 (0.4596)  | 7.7464 (0.5599)  | 8.4871 (0.5814)                            | 16.2054 (0.1337) | 16.7005 (0.1612) |  |  |
| OTAVA                     | 10.7944 (0.1478) | 11.0788 (0.1973) | 11.0893 (0.2696) | 11.1440 (0.3464)                           | 11.7296 (0.3843) | 11.7335 (0.4673) |  |  |
| PORUBKA                   | 6.7359 (0.4569)  | 10.6075 (0.2249) | 10.7152 (0.2957) | 11.3431 (0.3314)                           | 11.6329 (0.3919) | 12.0769 (0.4395) |  |  |
| ROZNOV.BECVA              | 4.9917 (0.6610)  | 6.5629 (0.5844)  | 10.9195 (0.2813) | 11.7738 (0.3005)                           | 12.2054 (0.3484) | 13.8705 (0.3091) |  |  |
| SVRATKA                   | 3.0177 (0.8834)  | 3.2377 (0.9186)  | 3.6206 (0.9346)  | 4.4708 (0.9236)                            | 4.4956 (0.9531)  | 5.9885 (0.9167)  |  |  |
| UPA                       | 6.9100 (0.4383)  | 7.0184 (0.5347)  | 7.0276 (0.6342)  | 9.4851 (0.4868)                            | 9.9573 (0.5342)  | 10.7439 (0.5510) |  |  |
| VLARA                     | 9.1310 (0.2434)  | 10.3181 (0.2434) | 10.5452 (0.3082) | 11.4597 (0.3228)                           | 11.6880 (0.3875) | 12.4812 (0.4078) |  |  |
| VYDRA                     | 8.3236 (0.3049)  | 11.2454 (0.1882) | 13.6099 (0.1369) | 13.8708 (0.1790)                           | 14.7466 (0.1944) | 16.0820 (0.1875) |  |  |
| ZDOBNICE                  | 5.1371 (0.6432)  | 6.8762 (0.5500)  | 7.0328 (0.6337)  | 10.9401 (0.3622)                           | 13.1479 (0.2838) | 17.4549 (0.1333) |  |  |

**Tab. 2** *Výsledné hodnoty při použití metody hlavních komponent*

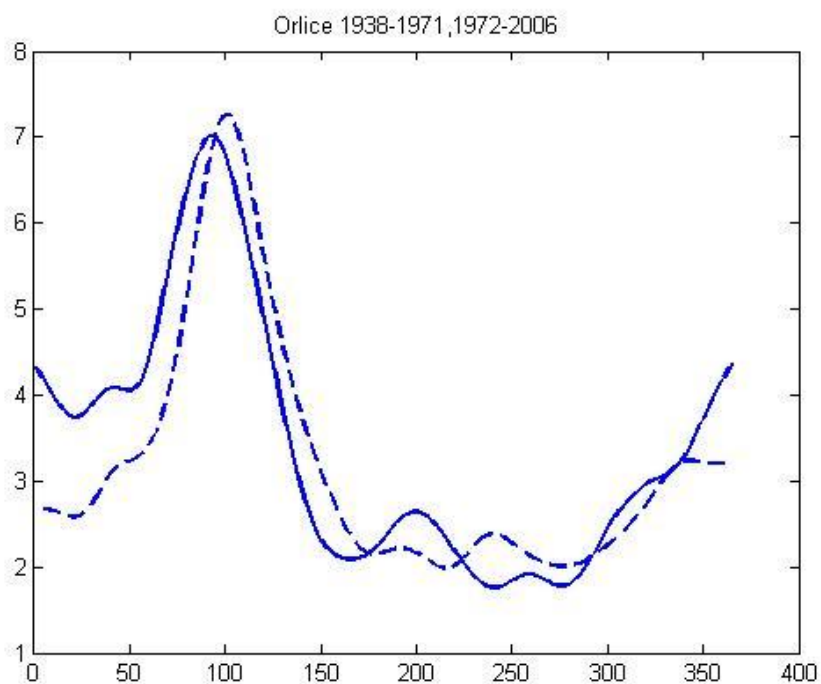


**Graf 1** Vyhlazený roční chod za období 1953-1980 (čárkovaná), 1981-2008 (plná)

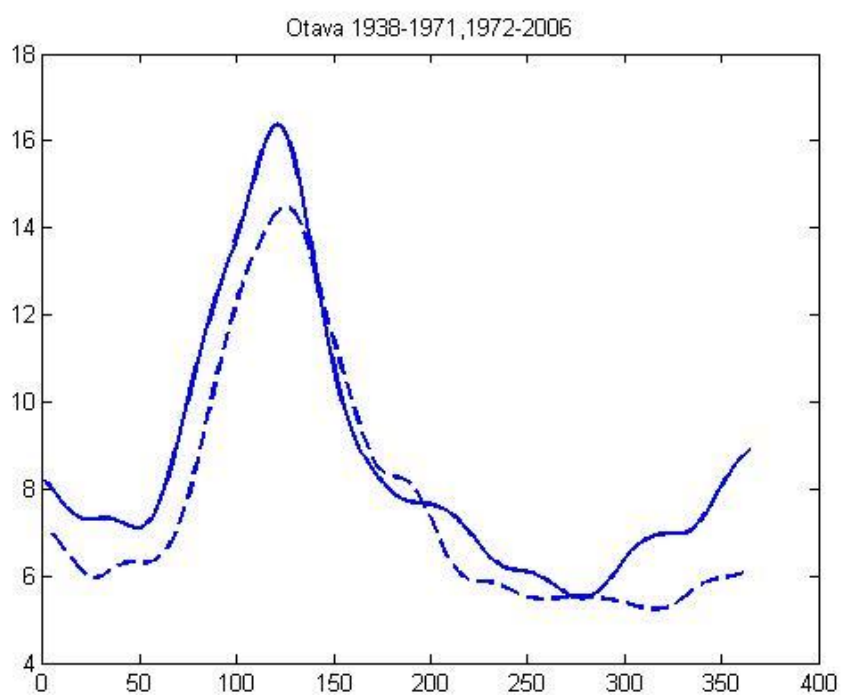


**Graf 2** Vyhlazený roční chod za období 1941-1974 (čárkovaná), 1975-2008 (plná)

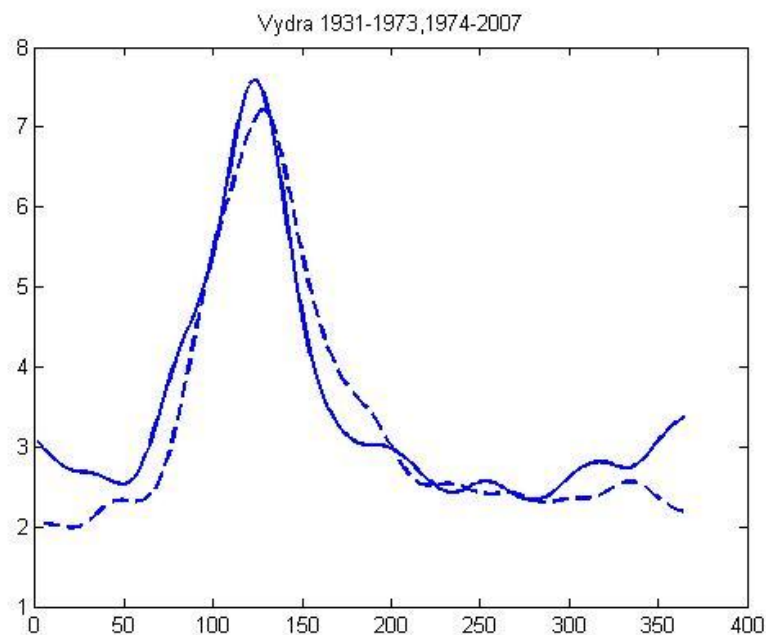




**Graf 3** Vyhlazený roční chod za období 1938-1971 (čárkovaná), 1972-2006 (plná)



**Graf 4** Vyhlazený roční chod za období 1938-1971 (čárkovaná), 1972-2006 (plná)



**Graf 5** Vyhlazený roční chod za období 1931-1973 (čárkovaná), 1974-2007 (plná)

## 6 Poděkování

Tato práce byla podpořena grantem studentské grantové soutěže ČVUT v Praze č. SGS12/004/OHK1/1T/11 „Stochastické vlastnosti denních průtokových dat“. Data pocházejí z projektu NAZV QH71201 „Spolehlivost a bezpečnost vodohospodářských děl v měnících se klimatických podmínkách“.

## Literatura

- [1] JARUŠKOVÁ, Daniela. *Pravděpodobnost a matematická statistika*. Praha: Česká technika - nakladatelství ČVUT, 2006. 138 s. ISBN 80-01-03427-5
- [2] ANDERSON, Theodore, Wilbur. *An introduction to multivariate statistical analysis*. New York: Wiley & Sons, 1971. 752 s. ISBN 0-471-36091-0
- [3] ANDERSON, Theodore, Wilbur. *The statistical analysis of time series*. New York: John Wiley & Sons. 1994. 704 s. ISBN 0-471-04745-7