Vyhledávání změn v průtokových řadách

Daniela Jarušková, stavební fakulta Praha

1. Úvod

Hydrologové na celém světě se obávají změn, které mohou nastat v hydrologických řadách, především v průtocích řek, díky silnému zasahování člověka do přírodních dějů během posledních desetiletí. Aby bylo možno případné změny zjistit a posoudit, je třeba zkoumat, zda i v minulosti nedocházelo v průběhu průtokových řad k určitým změnám. Délka naměřených průměrných ročních průtokových řad se v našich zemích pohybuje kolem 70, s výjimkou Labe, kde je známa průtoková řada o 138 údajích. Pro statistické účely je možno předpokládat, že se průměrné roční průtoky řídí tříparametrickým log-normálním rozdělením. To znamená, že transformací $Y_i = \ln(X_i + a)$, $i=1,\dots,n$, získáme normálně rozdělené náhodné veličiny, o kterých lze obvykle předpokládat, že jsou nezávislé nebo tvoří AR či ARMA posloupnost. Pro vyhledávání případných změn lze použít následujících postupů matematické statistiky.

2. Testování změn v modelu

Uvažujme následující hypotézy.

I. H_o: $X_1,\dots,X_n$ jsou nezávislé náhodné veličiny řídící se hustotou $f(x, \underline{\Theta}_o)$

A : $X_1,\dots,X_n$ jsou nezávislé náhodné veličiny,
existuje $k \in \{1,\dots,n-1\}$ takové, že
$X_1,\dots,X_k$ se řídí hustotou $f(x, \underline{\Theta}')$, $X_{k+1},\dots,X_n$ se řídí hustotou $f(x, \underline{\Theta}'')$,
přičemž $\underline{\Theta}' \neq \underline{\Theta}''$.

II. H_o: $X_1,\dots,X_n$ tvoří zobecněnou autoregresní posloupnost p-tého řádu
$AR_{(\mu_o, \sigma_o^2, \underline{a}_o)}(p)$ splňující:
$(X_t - \mu_o) = a_{o1}(X_{t-1} - \mu_o) + \dots + a_{op}(X_{t-p} - \mu_o) + e_t$, kde e_t jsou nez. náh. vel. řídící se $N(0, \sigma_o^2)$.

A : existuje $k \in \{1,\dots,n-1\}$ takové, že
$X_1,\dots,X_k$ tvoří zobecněnou autoregresní posloupnost $AR_{(\mu', \sigma'^2, \underline{a}')}(p)$,
$X_{k+1},\dots,X_n$ tvoří zobecněnou autoregresní posloupnost $AR_{(\mu'', \sigma''^2, \underline{a}'')}(p)$,
$\underline{\Theta}' = (\mu', \sigma'^2, \underline{a}')$, $\underline{\Theta}'' = (\mu'', \sigma''^2, \underline{a}'')$ a $\underline{\Theta}' \neq \underline{\Theta}''$.

III. H_o: $X_1,\dots,X_n$ jsou nezávislé náh.vel. řídící se rozdělením s absolutně spojitou distribuční funkcí $F(x)$

A : $X_1,\dots,X_n$ jsou nezávislé náh.vel.
existuje $k \in \{1,\dots,n-1\}$ takové, že
$X_1,\dots,X_k$ jsou nez.náh.vel.řídící se rozdělením s absolutně spojitou

distribuční funkcí $F'(x)$,

$X_{k+1},\dots,X_n$ jsou nez.náh.vel.řídící se rozdělením s absolutně spojitou distribuční funkcí $F''(x)$.

Řešení případů I a II

Testujeme-li nulovou hypotézu proti alternativě, že došlo ke změně v předem známém bodě k, je vhodnou testovou statistikou poměr věrohodnosti, přičemž se neznámé parametry nahradí jejich maximálně věrohodnými odhady:

$$\Lambda\left(\frac{k}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) = \sup_{\underline{\Theta}'}\sup_{\underline{\Theta}''}\inf_{\underline{\Theta}} \frac{f(x_1,\dots,x_k;\underline{\Theta}')\cdot f(x_{k+1},\dots,x_n;\underline{\Theta}'')}{f(x_1,\dots,x_n;\underline{\Theta})} .$$

Jestliže navíc bod k, v kterém došlo ke změně neznáme, je přirozené použít testové statistiky:

$$\sup_k \Lambda\left(\frac{k}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right).$$

Přesne rozdělení lze spočítat jen v některých velmi speciálních případech, a proto je třeba pro hledání kritických hodnot použít asymptotické rozdělení. Za jistých technických předpokladů lze ukázat, že za platnosti nulové hypotézy platí pro $\forall A \geq 1$

$$\lim_{n\to\infty} P\left(\sup_{1\leq k\leq n} \Lambda\left(\frac{k}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) > A\right) = 1,$$

což je způsobeno tím, že $\Lambda\left(\frac{k}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right)$ nabývají pro k blízká 0 a n velkých hodnot. Dále za platnosti nulové hypotézy platí pro $\forall \varepsilon < \frac{1}{2}$ a $\forall A \geq 1$

$$\lim_{n\to\infty} P\left(\sup_{n\varepsilon\leq k\leq n(1-\varepsilon)} \Lambda\left(\frac{k}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) > A\right) = P\left(\sup_{\varepsilon<t<(1-\varepsilon)} \Lambda\left(t, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) > A\right),$$

přičemž limitní proces je definován následovně:

$$\Lambda\left(t, \underline{\Theta}', \underline{\Theta}'', \underline{\Theta}\right) = \frac{L(t, \underline{\Theta}')\cdot L(1, \underline{\Theta}'')}{L(t, \underline{\Theta}'')\cdot L(1, \underline{\Theta})},$$

$$L(t, \underline{\Theta}) = \exp\left\{\underline{\Theta}^T\cdot\left[\underline{\underline{I}}(\Theta_0)\right]^{1/2}\cdot \underline{W}_t - \frac{t}{2}\underline{\Theta}^T\cdot\underline{\underline{I}}(\Theta_0)\cdot\underline{\Theta}\right\},$$

kde $\underline{\underline{I}}(\Theta_0)$ je Fisherova informační matice a $\underline{W}_t$ je vícerozměrný Wienerův proces. Odtud je patrné, že jakýmkoliv normováním spočívajícím v násobení věrohodnostního poměru konstantami $C(n)$ nelze získat rozumné asymptotické rozdělení. Možnost řešení spočívá buď ve vyloučení krajních bodů (t=0 a t=1) nebo ve vhodné penalizaci obou konců.

V modelu I předpokládejme, že hustota $f(x, \underline{\Theta})$ splňuje určité technické podmínky, viz [2]. Tyto podmínky jsou splněny např. pro normální nebo gamma rozdělení.

V modelu II předpokládejme, že kořeny charakteristického polynomu $z^p - a_1 z^{p-1} - \dots - a_p = 0$ leží uvnitř jednotkového kruhu.

Za těchto předpokladů pro logaritmický poměr věrohodnosti

$$\mathcal{L}(\tfrac{k}{n}, \underline{\theta}'^{*}, \underline{\theta}''^{*}, \underline{\theta}^{*}) = \sup_{\underline{\theta}'} \sup_{\underline{\theta}''} \inf_{\underline{\theta}} \ln \frac{f(x_1,\ldots,x_k;\underline{\theta}')\cdot f(x_{k+1},\ldots,x_n;\underline{\theta}'')}{f(x_1,\ldots,x_n;\underline{\theta})}$$

platí:

$$(1)\quad P\Big(\sup_{n\varepsilon \le k \le n(1-\varepsilon)} 2\,\mathcal{L}(\tfrac{k}{n}, \underline{\theta}'^{*}, \underline{\theta}''^{*}, \underline{\theta}^{*}) > v\Big) \longrightarrow P\Big(\sup_{\varepsilon \le t \le 1-\varepsilon} \frac{\sum_{i=1}^{p} B_i^2(t)}{t(1-t)} > v\Big)$$

$$(2)\quad P\Big(\sup_{1 \le k \le n} 2\tfrac{k}{n}(1-\tfrac{k}{n})\,\mathcal{L}(\tfrac{k}{n}, \underline{\theta}'^{*}, \underline{\theta}''^{*}, \underline{\theta}^{*}) > v\Big) \longrightarrow P\Big(\sup_{0<t<1} \sum_{i=1}^{p} B_i^2(t) > v\Big)\ .$$

Proces $\sum_{i=1}^{p} B_i^2(t)$ je součet kvadrátů nezávislých Brownových můstků, přičemž p je dimenze parametru $\underline{\Theta}$. Distribuční funkce $\sup_{0<t<1} \sum_{i=1}^{p} B_i^2(t)$ byla tabelována např. v [6], kritické hodnoty této statistiky byly uvedeny v [4]. Pro aproximaci kritických hodnot statistiky $\sup_{\varepsilon<t<(1-\varepsilon)} \sum B_i^2(t) / t(1-t)$ lze použít vztah:

$$(3)\quad P\Big(\sup_{t_0 \le t \le t_1} \sqrt{\textstyle\sum B_i^2(t) / t(1-t)} > b\Big) = \frac{b^p \exp(-\frac{b^2}{2})}{2^{(p-2)/2}\cdot\Gamma(\frac{p}{2})}\Big(\frac{(1-p/b^2)}{2}\cdot \log r + \frac{2}{b^2}\Big)$$

kde $r = t_1(1-t_0)/t_0(1-t_1)$, viz [5].

Poznamenejme, že v případě II lze pro velká n maximálně věrohodný odhad nahradit odhadem podmíněně maximálně věrohodným. Logaritmický věrohodnostní poměr má pak pro případ normálního rozdělení pro I a II tvar

$$\mathcal{L}(\tfrac{k}{n}, \theta'^{*}, \theta''^{*}, \theta^{*}) = (n/2)\ln \hat{\sigma}^2 - (k/2)\ln \hat{\sigma}'^2 - ((n-k)/2)\ln \hat{\sigma}''^2\ ,$$

kde $\hat{\sigma}^2$, $\hat{\sigma}'^2$, $\hat{\sigma}''^2$ jsou odhady příslušných rozptylů.

Řešení případu III

Pro testování nulové hypotézy v modelu III je možné použít penalizovanou Kolmogorov-Smirnovovu statistiku, viz [3] nebo [4].

$$\sup_k \sup_x \tfrac{k}{n}(1-\tfrac{k}{n})\sqrt{n}\,|\hat{F}'(x) - \hat{F}''(x)|\ .$$

Asymptotickou kritickou hodnotu můžeme odvodit z limitního chování dané statistiky:

$$(4)\quad P\Big(\sup_k \sup_x \tfrac{k}{n}(1-\tfrac{k}{n})\sqrt{n}\,|\hat{F}'(x) - \hat{F}''(x)| > v\Big) \longrightarrow P\Big(\sup_{t\in(0,1)} \sup_{u\in(0,1)} |B(u,t)| > v\Big),$$

kde B(u,v) je Brownův list.

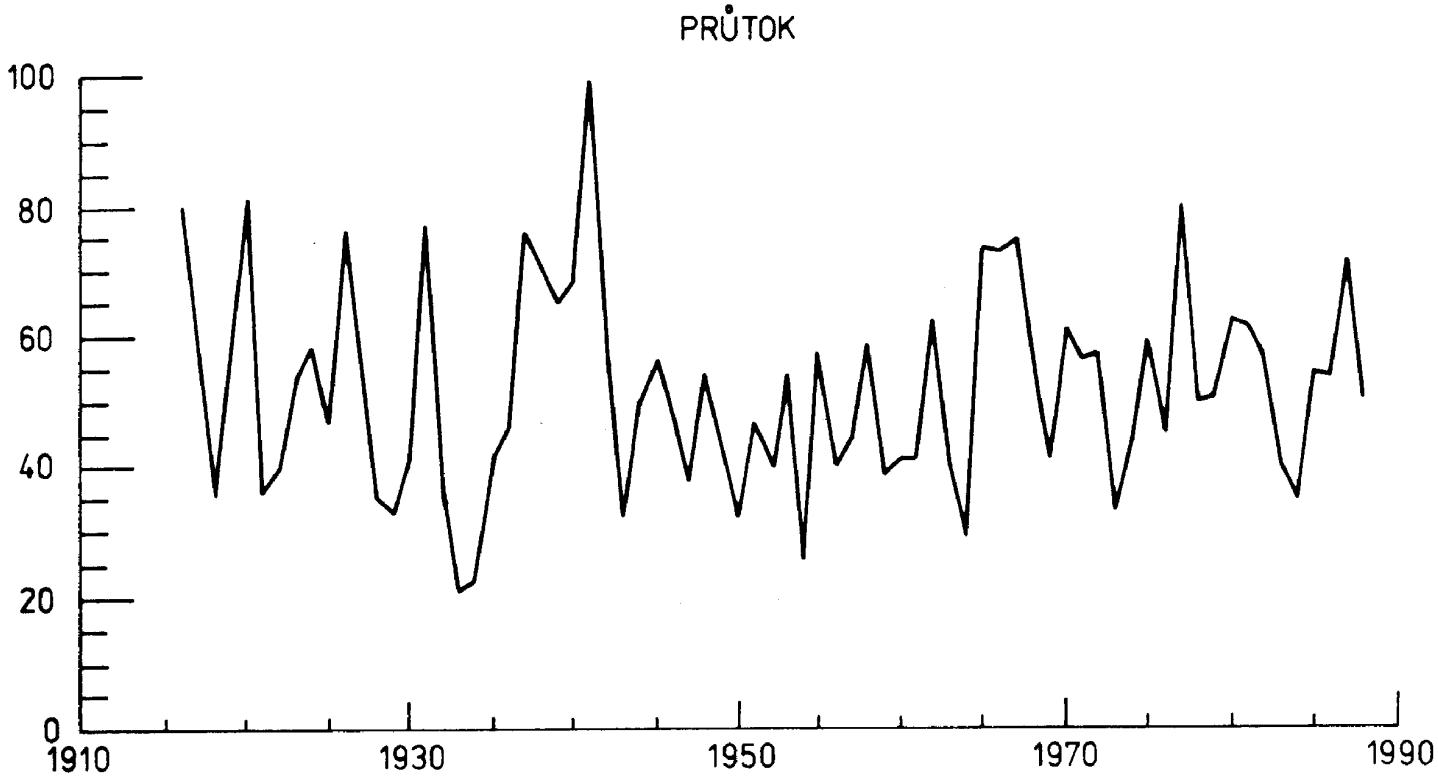
3. Použití testů pro hledání změn v ročních průtokových řadách

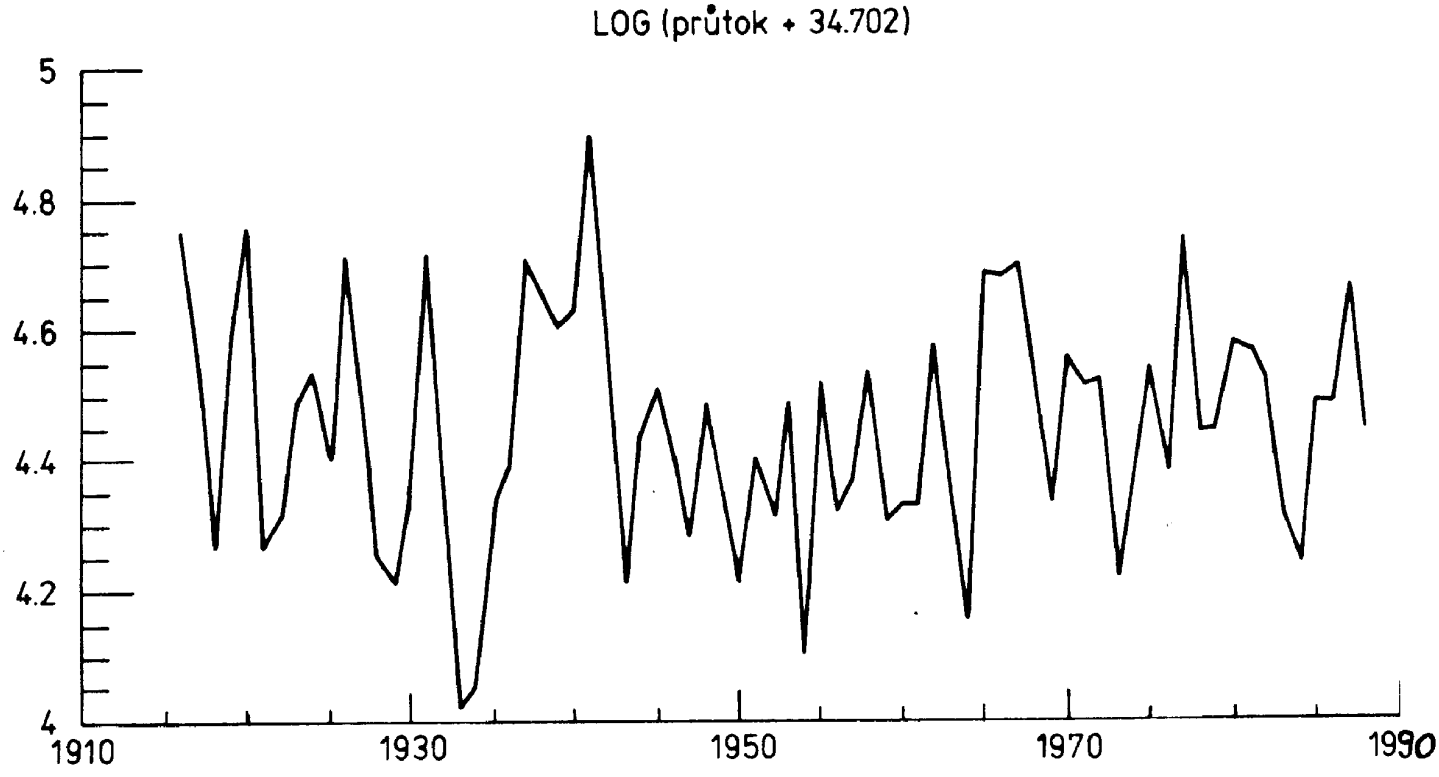
Pro každou sledovanou řeku jsme uvažovali roční průtokovou řadu $X_1,\ldots,X_n$ a transformovanou řadu $Y_1 = \ln (X_1 + a),\ldots,Y_n = \ln (X_n + a)$, kde a bylo voleno tak, aby bylo možno předpokládat, že $Y_1,\ldots,Y_n$ jsou normálně rozdělené. Model I jsme použili pro řadu $Y_1,\ldots,Y_n$, kde změna modelu spočívala ve změně parametru $\underline{\theta} = (\mu,\sigma^2)$. V modelu II jsme předpokládali, že $Y_1,\ldots,Y_n$ tvoří zobecněnou autoregresní posloupnost 1.řádu. Změna modelu spočívala ve změně parametru $\underline{\theta} = (\mu,\sigma^2,a)$. Model III jsme použili pro původní řadu $X_1,\ldots,X_n$.

Příklad

V tabulce jsou uvedeny průměrné roční průtoky řeky Moravy měřené v Kroměříži a hodnoty získané transformací $y = \ln (x + 34.702)$.

poř.	rok	X_i	Y_i	poř.	rok	X_i	Y_i	poř.	rok	X_i	Y_i
1	1916	80.069	4.7429388	25	1940	68.154	4.6333300	49	1964	29.347	4.1596484
2	1917	62.921	4.5811131	26	1941	99.027	4.8958154	50	1965	73.278	4.6819460
3	1918	36.327	4.2630882	27	1942	56.661	4.5148406	51	1966	72.874	4.6781976
4	1919	64.701	4.5991823	28	1943	32.469	4.2072416	52	1967	74.654	4.6946086
5	1920	80.877	4.7499543	29	1944	49.515	4.4333968	53	1968	57.078	4.5193944
6	1921	36.281	4.2624404	30	1945	56.267	4.5105188	54	1969	41.616	4.3349088
7	1922	40.472	4.3198054	31	1946	48.285	4.4186840	55	1970	60.676	4.5578479
8	1923	53.844	4.4835222	32	1947	37.695	4.2821649	56	1971	56.425	4.5122541
9	1924	58.101	4.5304790	33	1948	54.308	4.4887487	57	1972	57.178	4.5204834
10	1925	47.159	4.4050227	34	1949	41.637	4.3351839	58	1973	33.462	4.2219166
11	1926	76.113	4.7078621	35	1950	32.600	4.2091900	59	1974	43.304	4.3567857
12	1927	57.710	4.5262568	36	1951	46.892	4.4017557	60	1975	58.910	4.5391586
13	1928	35.570	4.2523734	37	1952	40.194	4.3161005	61	1976	45.049	4.3789093
14	1929	32.776	4.2118016	38	1953	54.027	4.4855868	62	1977	79.585	4.7387128
15	1930	42.078	4.3409442	39	1954	25.681	4.1007076	63	1978	50.009	4.4392455
16	1931	77.058	4.7163537	40	1955	57.151	4.5201895	64	1979	50.387	4.4436978
17	1932	35.418	4.2502081	41	1956	40.317	4.3177414	65	1980	62.232	4.5740303
18	1933	21.022	4.0204109	42	1957	44.254	4.3688907	66	1981	61.595	4.5674372
19	1934	22.936	4.0541821	43	1958	58.401	4.5337064	67	1982	57.377	4.5226469
20	1935	41.612	4.3348564	44	1959	39.206	4.3028211	68	1983	39.954	4.3128909
21	1936	46.399	4.3956953	45	1960	41.169	4.3290345	69	1984	34.588	4.2383006
22	1937	75.933	4.7062365	46	1961	41.100	4.3281247	70	1985	54.112	4.4865443
23	1938	69.522	4.6465424	47	1962	62.269	4.5744120	71	1986	53.885	4.4839851
24	1939	65.180	4.6039895	48	1963	39.788	4.3106649	72	1987	71.098	4.6615505
								73	1988	50.582	4.4459869





Model I

$\underline{\theta} = (\mu, \sigma^2)$, p = 2

$k^* = 25$

	1916 - 1941	1942 - 1988
$\hat{\mu}$	$= \bar{y} = 4.471$	$= \bar{y} = 4.433$
$\hat{\sigma}^2$	$= s_y^2 = 0.050$	$= s_y^2 = 0.022$

$$\mathcal{L}\left(\frac{k^*}{n}, \Theta'^*, \Theta''^*, \Theta^*\right) = 3.551$$

$$2\frac{k^*}{n}\left(\frac{n-k^*}{n}\right)\mathcal{L}\left(\frac{k^*}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) = 1.629$$

Použijeme-li aproximace (2), pak p-hodnota odpovídající příslušné penalizované statistice je rovna 0.28 .

Model II

$\underline{\Theta} = (\mu, \sigma^2, a)$, $p = 3$ | 1916 - 1940 | 1941 - 1988

$k^* = 25$

$\hat{\mu}' = 4.439$ | $\hat{\mu}'' = 4.432$

$\hat{a}' = 0.3486$ | $\hat{a}'' = 0.0712$

$\hat{\sigma}^2 = 0.03759$ | $\hat{\sigma}''^2 = 0.02141$

$$\mathcal{L}\left(\frac{k^*}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) = 4.6994$$

$$2\frac{k^*}{n}\frac{(n-k^*)}{n}\mathcal{L}\left(\frac{k^*}{n}, \underline{\Theta}'^*, \underline{\Theta}''^*, \underline{\Theta}^*\right) = 2.116451$$

Použijeme-li aproximace (2), pak p-hodnota odpovídající příslušné penalizované statistice je rovna 0.22 .

Model III

Penalizovaná Kolmogorov-Smirnovova statistika nabývá maxima pro $k^* = 26$ (rok 1941)

$$\sup_x \frac{k^*}{n}\left(1 - \frac{k^*}{n}\right)\sqrt{n}\left|\hat{F}'(x) - \hat{F}''(x)\right| = 0.620,$$

čemuž, použijeme-li aproximace (4), odpovídá p-hodnota 0.26 .

Závěr

Všechny postupy se shodly v tom, že k případné změně by mohlo dojít nejspíš v roce 1940 či 1941 ($k^* = 25,26$). Z obrázků 1,2 se skutečně zdá, že v tomto období došlo ke změně v charakteru průběhu řady - změnila se rozptýlenost řady i závislost mezi následujícími členy řady. Z hlediska statistiky však změna nebyla významná, neboť p-hodnota byla ve všech případech daleko větší než 0.05 . Pro zajímavost uveďme, že druhým případným bodem změny (ovšem ještě slaběji prokazatelným) byl určen rok 1964.

4. Hodnocení uvedených metod

Zdůrazněme znovu, že uvedené metody slouží k odhalení jedné náhlé změny v modelu. Jestliže změna probíhá pozvolna nebo dochází k více než jedné změně, popsané metody dávají horší výsledky. Přesto uveďme, že ve všech našich případech spolehlivě vyhledávaly body změny tam, kde bylo z obrázku patrné, že k určité změně v průběhu řady dochází. Odpovídající p-hodnoty získané aproximací (1) a (2) však doporučujeme brát jen informačně. Rychlost konvergence k asymptotickému rozdělení není známa a bude patrně velmi záviset na skutečném rozdělení pozorovaných veličin.

LITERATURA

[1] Deshayes,J.,Picard,D.(1984). Principe d´invariance sur le processus de vraisemblance. Annals de l´I.H.P., vol.20, pp.1-20

[2] Deshayes,J.,Picard,D.(1984). Lois asymptotiques des tests et estimateurs de rupture dans un modele statistique classique. Annals de l´I.H.P., vol. 20, pp. 309-327

[3] Deshayes,J.,Picard,D.(1986). Off-line statistical analysis of change-point models. Lecture Notes in Control and Information Sciences 77 - Detection of abrupt changes in signals and dynamical systems, pp.103-168

[4] Hušková,M.(1988). Detekce změny regrese a detekce změny rozdělení. Sborník ROBUST 88.

[5] James,B.,James,K.L.,Siegmund,D.(1987). Tests for a change-point. Biometrika 74, pp.71-83

[6] Kiefer,J.(1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér- - V.Mises tests, AMS 30, pp.420-447

[7] Picard,D.(1985). Testing and estimating change-points in time series, Adv. Appl. Prob.17, pp.841-867