Pokroky matematiky, fyziky a astronomie

Daniela Jarušková; Jaromír Antoch

Detekce změn v časových řadách a její aplikace v ekologii

Pokroky matematiky, fyziky a astronomie, Vol. 47 (2002), No. 4, 307--323

Persistent URL: http://dml.cz/dmlcz/141146

Detekce změn v časových řadách a její aplikace v ekologii

Daniela Jarušková a Jaromír Antoch, Praha

Abstrakt

V přírodních vědách, ekonomii i v technice se často objevují problémy spojené se zjišťováním nestacionárního chování časových řad. Je přitom přirozené, že k tomuto studiu se používají metody matematické statistiky. Na druhé straně je pro řadu uživatelů překvapením, že k danému účelu zpravidla nevystačí ani s klasickými metodami lineární a nelineární regrese, ani se standardními postupy analýzy časových řad. Naopak odhalování a odhadování bodu změny vyžaduje většinou zvláštní aparát. Čtenář může být udiven, že jednoduše formulované problémy, jejichž řešení každý „vidí", vyžadují v řadě případů (velmi) složitou teorii.

Cílem tohoto článku je čtenáři na několika příkladech z meteorologie, hydrologie a ekologie předvést, jak lze v rámci matematické statistiky zformulovat problém týkající se zjišťování nestacionárního chování časové řady pomocí teorie testování hypotéz. Pro odvození testových statistik jsou popsány dvě základní metody, a to metoda maximální věrohodnosti a metoda pseudo-bayesovská, spolu s postupy umožňujícími aproximovat příslušné kritické hodnoty. Pokud nulovou hypotézu zamítneme, tj. přikloníme se k názoru, že ke změně modelu došlo, následuje druhý krok statistické inference — určení času, kdy ke změně došlo. Takový časový okamžik se nazývá *bod změny* (anglicky *change point*). K odhadu bodu změny, stejně jako ostatních parametrů modelu, které například odpovídají velikosti změny, k níž došlo, se používá statistická teorie odhadu. Uvedené postupy jsou nakonec použity k vyšetřování, zda mohou být časové řady, s nimiž se čtenář seznámil v úvodu článku, považovány za stacionární, případně kde a k jak velké změně došlo.

1. Úvod

V meteorologii, klimatologii, hydrologii a studiích o životním prostředí se poměrně dlouhou dobu pravidelně zaznamenávají veličiny jako teplota, tlak, množství atmosférického ozónu, průtoky vody daným profilem apod. Pozorované časové řady přitom zcela přirozeně vykazují náhodné fluktuace, a tudíž jejich chování může být popsáno stochastickými modely. Při studiu těchto řad vyvstává důležitá otázka, zda pozorované časové řady jsou „stabilní" v čase, tj. zda pozorované řady jsou *stacionární*, či zda v nějakém okamžiku (období), zpravidla pozorovateli neznámém, změnily své chování. Zdůrazněme, že termín „stacionární chování časové řady" zde pokrývá mnohem širší

Doc. RNDr. Daniela Jarušková, CSc. (1951), ČVUT v Praze, Stavební fakulta, katedra matematiky, Thákurova 7, 166 29 Praha 6; doc. RNDr. Jaromír Antoch, CSc. (1953), MFF UK, katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 186 75 Praha 8.

situaci, než statistikové pod tímto pojmem většinou chápou. Pro potřeby tohoto příspěvku považujeme časovou řadu za stacionární například i tehdy, když její průměrný přírůstek zůstává týž v čase apod. Ze statistického hlediska se změna v časové řadě vyskytne, jestliže existuje takový časový moment, že chování řady do tohoto okamžiku lze popsat jedním pravděpodobnostním modelem, zatímco po něm modelem jiným.

V praxi jsou samozřejmě některé z těchto změn způsobeny „přirozenou variabilitou". Tak například je známo, že se v minulosti střídala období s chladným počasím (tzv. doby ledové) s obdobími teplejšími. V jiných případech mohou být změny způsobeny lidskou činností, jako je například odlesňování, zvyšování spotřeby fosilních paliv, přesun lidí do velkých měst, nadměrné používání automobilů atd. Toto vše vede řadu badatelů k diskusím o tom, zda nedochází k rozsáhlejším změnám klíčových klimatických veličin. Některé „změny" mohou být naproti tomu způsobeny výměnou měřicích přístrojů, změnou metodiky měření, změnou celkové situace v okolí měřicího bodu apod.

Třída modelů používaných v oblasti detekce strukturálních změn je velice široká. Změny se mohou vyskytovat v parametru polohy nebo měřítka, regresních koeficientech, korelační struktuře apod. Navíc změna může být ne jen jedna, ale může jich nastat i více. Odhadování počtu změn je velmi obtížný statistický problém, který nebyl dosud zcela vyřešen.

Statistická inference se zpravidla skládá ze dvou kroků. Nejprve je třeba rozhodnout, zda ke změně vůbec došlo. Rozhodnutí je založeno na testování hypotéz, kde nulová hypotéza tvrdí, že řada je stacionární, tj. k popisu řady lze použít jeden stochastický model, zatímco alternativa tvrdí, že existuje časový okamžik — bod změny — takový, že řadu před bodem změny lze popsat jedním modelem a po bodu změny jiným modelem. Uživatel tedy musí mít předem představu, jaký typ změny může v dané situaci nastat. Pokud byla změna detekována, je v druhém kroku třeba nalézt čas, v němž nastala. Nalezení okamžiku (či okamžiků) změny je úlohou teorie odhadu. Kromě bodového a intervalového odhadu bodu změny nás může rovněž zajímat odhad parametrů modelu před změnou i po ní, které mohou například sloužit k posouzení toho, k jak velké změně došlo.

Pro ilustraci metod detekce bodů změny ve stochastických modelech uveďme několik konkrétních příkladů týkajících se analýzy životního prostředí, které jsme v minulosti studovali.

Příklad 1.

Nehomogenity v časové řadě způsobené nekontrolovanými změnami měřicího procesu se často zjišťují porovnáním této řady s referenční časovou řadou, o které se předpokládá, že se v ní žádné nehomogenity nevyskytují, a která se chová podobně jako studovaná řada. Základními údaji v našem příkladu jsou dvě časové řady měsíčních průměrů atmosférického tlaku spočtené z denních měření ve dvou nepříliš vzdálených švýcarských meteorologických stanicích — Säntis a Gütsch. Problém spočívá ve zjištění případných nehomogenit ve stanici Gütsch, přičemž řada měsíčních průměrů ve stanici Säntis slouží jako referenční řada. Jestliže ve stanici Gütsch nedošlo ke změnám

měřicího procesu, pak by se vzhledem k rozdílné nadmořské výšce stanic měly obě řady lišit, až na malé náhodné odchylky, o konstantu.

Statistická analýza byla založena na $n = 240$ pozorovaných rozdílech mezi řadami naměřenými na jednotlivých stanicích (po odstranění sezónnosti). Předpokládali jsme přitom, že změna měřicího procesu mohla způsobit náhlou změnu ve střední hodnotě uvažovaných diferencí měření, o které předem nelze říci, zda může být kladná nebo záporná. Cílem statistické analýzy bylo rozhodnout, zda k takovému posunu došlo, a v případě kladné odpovědi odhadnout čas, v němž ke změně došlo, společně s velikostí této změny.

Podívejme se na celou otázku poněkud formálněji a označme $Y_1, \ldots, Y_n$ studovanou časovou řadu, v našem případě řadu sledovaných diferencí. Nulová hypotéza H_1 a alternativní hypotéza A_1 zde mohou být stanoveny následovně:

$$H_1 : \; Y_i = \mu + e_i, \quad i = 1, \ldots, n, \tag{1}$$

$$A_1 : \; \exists\, \mathsf{m} \in \{1, \ldots, n-1\} \;\text{ takové, že }\; \begin{cases} Y_i = \mu + e_i, & i = 1, \ldots, \mathsf{m}, \\ Y_i = \mu + \delta + e_i, & i = \mathsf{m}+1, \ldots, n. \end{cases}$$

Proměnné $\{e_i\}$ jsou náhodné chyby, $\delta \neq 0$ a $\mu \in \mathbb{R}^1$.

Příklad 2.

Mnoho klimatologů se domnívá, že zemské klima je ohroženo globálním oteplováním. Pro účely studia efektů globálního oteplování byla vytvořena řada „umělých časových řad“, blíže viz Jones et al. [22]. Autoři těchto řad si kladou za cíl popsat vývoj „globální zemské teploty“ kombinací teplotních řad získaných na mnoha meteorologických stanicích po celé zeměkouli. Jednou z nejznámějších teplotních řad používaných pro tyto účely je tzv. *series of global world temperature anomalies*, kterou sestavil Jones et al. [22]. Tato uměle zkonstruovaná řada začíná v roce 1854 a skládá se ze 140 hodnot, přičemž jedna hodnota odpovídá jednomu roku. Cílem statistické inference je zde rozhodnout, zda je možno Jonesovu řadu považovat za stacionární. Očekávaná změna přitom má tvar postupného růstu v průměru studované řady. Jestliže předpokládáme, že v neznámém časovém okamžiku došlo ke vzniku spojitého lineárního trendu (což je nejjednodušší typ postupné změny), je možné nulovou hypotézu H_2 a alternativu A_2 zapsat následovně:

$$H_2 : \; Y_i = a + e_i, \quad i = 1, \ldots, n, \tag{2}$$

$$A_2 : \; \exists\, \mathsf{m} \in \{1, \ldots, n-1\} \;\text{ takové, že }\; \begin{cases} Y_i = a + e_i, & i = 1, \ldots, \mathsf{m}, \\ Y_i = a + b \cdot \dfrac{i - \mathsf{m}}{n} + e_i, & i = \mathsf{m}+1, \ldots, n, \end{cases}$$

$a \in \mathbb{R}^1$, $b \neq 0$ a $\{e_i\}$ jsou náhodné chyby.

Příklad 3.

Rozsáhlá odlesňování, jichž jsme v celém světě každodenními svědky, mohou způsobit, že půda ztrácí svou schopnost zadržovat vodu. Pracovníci Výzkumného ústavu lesního hospodářství a myslivosti v Jílovišti-Strnadech studovali vliv kontrolovaného odlesňování na vztah mezi srážkami a odtoky. Cílem statistické analýzy bylo rozhodnout, zda se vztah mezi srážkami a odtoky v důsledku odlesňování mění. Pro zjednodušení modelu předpokládejme, že vztah mezi srážkami a odtoky je lineární. Úloha patří mezi mnoho jiných problémů detekce změny v modelu lineární regrese. Lze si položit otázku, zda došlo ke změně pouze v parametru posunutí nebo v parametru sklonu, resp. v obou parametrech, tj. v posunutí a/nebo sklonu. Problém detekce změny v alespoň jednom z regresních parametrů můžeme řešit testováním hypotéz:

$$H_3 : \; Y_i = a + bx_i + e_i, \quad i = 1, \ldots, n, \tag{3}$$

$$A_3 : \; \exists\, \mathsf{m} \in \{2, \ldots, n-2\} \;\text{ takové, že }\; \begin{cases} Y_i = a + bx_i + e_i, & i = 1, \ldots, \mathsf{m}, \\ Y_i = a_0 + b_0 x_i + e_i, & i = \mathsf{m}+1, \ldots, n, \end{cases}$$

kde $\{e_i\}$ jsou náhodné chyby, $a \neq a_0$ a/nebo $b \neq b_0$. Pro rozhodnutí, zda došlo ke změně v parametrech lineární závislosti mezi srážkami a průtoky, jsme měli k dispozici 36 dvojic dat ročních úhrnů srážek v dané oblasti a odpovídajících odtoků potoku Malá Ráztoka v Beskydech.

2. Jak rozhodnout, zda řada je stacionární

Jak jsme již ukázali, rozhodnutí, zda ke změně došlo, nebo ne, je možné založit na výsledku statistického testování hypotéz. Vlastní rozhodnutí, zda zamítnout nulovou hypotézu, jež tvrdí, že ke změně nedošlo, závisí na hodnotě testové statistiky. Ke konstrukci testových statistik se zde zpravidla používá buď metoda maximální věrohodnosti, nebo pseudo-bayesovská metoda. Na následujícím jednoduchém příkladu si ukažme, jak lze tyto postupy použít.

Příklad 4.

Nechť tatáž veličina je měřena pomocí dvou různých měřicích zařízení. Jelikož měření je ovlivněno náhodnými chybami, získané hodnoty se často liší. Předpokládejme, že na počátku procesu měření ani jedno z měřicích zařízení nemá systematickou chybu, tj. rozdíly mezi měřeními $\{Y_i\}$ jsou způsobeny pouze náhodnými chybami, a tudíž kolísají okolo nuly. Nicméně například vlivem závady na jednom z měřicích zařízení nebo změny podmínek, za jakých se měření provádí, může dojít k tomu, že se po nějakém čase rozdíly mezi měřeními začnou pohybovat kolem nějaké nenulové hodnoty μ.

Všimněme si, že tento příklad je velmi podobný příkladu 1. Rozdíl spočívá v tom, že v příkladu 1 nebyly měřicí přístroje umístěny na témže místě, a proto nebylo možno předpokládat, že rozdíly naměřených hodnot kolísají kolem nuly, ale museli jsme předpokládat, že kolísají kolem nějaké neznámé hodnoty, kterou bylo třeba odhadnout.

Vraťme se zpět k příkladu 4 a předpokládejme navíc pro větší jednoduchost a průhlednost výkladu, že se rozptyl studovaných diferencí během doby sledování nezměnil. Potom je naší úlohou testovat hypotézu H_4 proti alternativě A_4:

$$H_4: \ Y_i = e_i, \quad i = 1, \dots, n, \tag{4}$$

$$A_4: \ \exists\, \mathsf{m} \in \{0, \dots, n-1\} \ \text{ takové, že } \begin{cases} Y_i = e_i, & i = 1, \dots, \mathsf{m}, \\ Y_i = \mu + e_i, & i = \mathsf{m}+1, \dots, n, \end{cases}$$

kde $\mu \neq 0$. Někdy má smysl uvažovat místo oboustranné alternativy pouze alternativu jednostrannou, tj. $\mu > 0$ nebo $\mu < 0$. Je tomu tak tehdy, je-li znaménko případného posunu známo předem. Jelikož v našem příkladu jsou náhodné chyby $\{e_i\}$ chybami měření, můžeme předpokládat, že jsou nezávislé a normálně rozdělené. Pro jednoduchost navíc předpokládejme, že jejich rozptyl je roven jedné. Pokud je rozptyl známý a není roven nule, můžeme standardizací přejít k veličinám, které mají jednotkový rozptyl. Lze tedy předpokládat, že náhodné veličiny $\{e_i\}$ jsou nezávislé, stejně rozdělené a řídí se standardním normálním rozdělením $N(0,1)$ s hustotou $\varphi(x)$ a distribuční funkcí $\Phi(x)$, tj. pro $x \in \mathbb{R}^1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} \quad \text{a} \quad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\, \mathrm{d}t.$$

Metoda maximální věrohodnosti

Metoda maximální věrohodnosti je jedním z nejpoužívanějších postupů při testování hypotéz. Její hlavní myšlenkou je použití věrohodnostního poměru, případně statistiky s ním ekvivalentní k testování nulové hypotézy proti alternativě. Podrobnější vysvětlení této metody nalezne čtenář v knize Hátleho a Likeše [15].

Jestliže je $\mu \neq 0$, potom logaritmus věrohodnostního poměru pro testování H_4 proti A_4 má tvar

$$\max_{0 \leqq k \leqq n-1} \sup_{\mu} \log \frac{\prod_{i=1}^{k} \varphi(Y_i) \prod_{i=k+1}^{n} \varphi(Y_i - \mu)}{\prod_{i=1}^{n} \varphi(Y_i)} = \max_{0 \leqq k \leqq n-1} \frac{1}{2(n-k)} \Big(\sum_{i=k+1}^{n} Y_i \Big)^2,$$

takže pro oboustrannou alternativu $\mu \neq 0$ má testová statistika, která je ekvivalentní s poměrem věrohodnosti, tvar:

$$\max_{0 \leqq k \leqq n-1} \left\{ \left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right| \right\}. \tag{5}$$

Pro jednostrannou alternativu s $\mu > 0$ lze obdobně odvodit testovou statistiku ve tvaru

$$\max_{0 \leqq k \leqq n-1} \left\{ \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right\}. \tag{6}$$

Tyto statistiky patří mezi tzv. *testové statistiky maximálního typu*. Velké hodnoty testové statistiky vedou k rozhodnutí nulovou hypotézu zamítnout. Více se o principu testování statistických hypotéz lze dozvědět například z knih Anděl [3] nebo Hátle a Likeš [15].

Pseudo-bayesovská metoda

Metoda je založena na předpokladu, že neznámý bod změny m a neznámá velikost posunu μ jsou náhodné veličiny se známým apriorním rozdělením. Velmi často se předpokládá, že apriorní rozdělení m je rovnoměrné, tj. $P(\mathsf{m}=i)=1/n$, $i=0,\dots,n-1$, a μ se řídí normálním rozdělením $N(0,\gamma^2)$. Za předpokladu, že hustota $Y_1,\dots,Y_n$ při daném $\mu=\mu$ a $\mathsf{m}=k$ je také normální, dostáváme

$$f\big(y_1,\dots,y_n \mid \mu=\mu,\, \mathsf{m}=k\big)=\prod_{i=1}^{k}\varphi(y_i)\prod_{i=k+1}^{n}\varphi(y_i-\mu),$$

takže nepodmíněná hustota může být vyjádřena ve tvaru

$$\begin{aligned} f(y_1,\dots,y_n) &= \sum_{k=1}^{n}\frac{1}{n}\int_{-\infty}^{\infty}\prod_{i=1}^{k}\varphi(y_i)\prod_{i=k+1}^{n}\varphi(y_i-\mu)\,\frac{1}{\gamma\sqrt{2\pi}}\exp\big\{-\mu^2/2\gamma^2\big\}\,\mathrm{d}\mu = \\ &= \prod_{i=1}^{n}\varphi(y_i)\left(\sum_{k=1}^{n}\frac{1}{n}\int_{-\infty}^{\infty}\frac{1}{\gamma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(-2\mu\sum_{i=k+1}^{n}y_i+(n-k)\mu^2+\frac{\mu^2}{\gamma^2}\right)\right\}\,\mathrm{d}\mu\right) \end{aligned}$$

a odpovídající věrohodnostní poměr má tvar

$$\frac{f_A(Y_1,\dots,Y_n)}{f_H(Y_1,\dots,Y_n)}=\sum_{k=1}^{n}\frac{1}{n}\sqrt{\frac{1}{1+(n-k)\gamma^2}}\,\exp\left\{\frac{\gamma^2}{2\big(1+(n-k)\gamma\big)^2}\Big(\sum_{i=k+1}^{n}Y_i\Big)^2\right\}.$$

Použijeme-li limitní přechod $\gamma\to 0$ a aplikujeme-li Taylorův rozvoj, je testová statistika, kterou dostaneme při použití věrohodnostního poměru pro oboustrannou alternativu, ekvivalentní testové statistice

$$\sum_{k=1}^{n}\frac{1}{n}\Big(\frac{1}{\sqrt{n}}\sum_{i=k+1}^{n}Y_i\Big)^2. \tag{7}$$

Získaná statistika patří mezi tzv. *testové statistiky součtového typu*.

Pro jednostrannou alternativu s $\mu>0$ můžeme analogicky odvodit testové statistiky součtového typu tvaru

$$\sum_{k=1}^{n}\frac{1}{n}\Big(\frac{1}{\sqrt{n}}\sum_{i=k+1}^{n}Y_i\Big)=\frac{1}{n}\,\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(k-1)Y_k. \tag{8}$$

Pseudo-bayesovská metoda byla navržena Kanderem a Zachsem [23].

Kritické hodnoty

Pro rozhodnutí, zda zamítnout nulovou hypotézu, potřebujeme kritické hodnoty použité testové statistiky. Pro jejich nalezení musíme znát rozdělení odpovídajících testových statistik za platnosti nulové hypotézy.

Začněme nejprve s testovou statistikou (5). Pokud jsou veličiny $\{Y_i\}$ nezávislé, stejně rozdělené a řídí se standardním normálním rozdělením $N(0,1)$, potom statistiky

$$\frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i, \quad k = 0, \ldots, n-1, \tag{9}$$

mají též standardní normální rozdělení $N(0,1)$. Abychom stanovili přesné rozdělení statistiky (5), je třeba nalézt rozdělení maxima absolutních hodnot náhodných veličin se standardním normálním rozdělením, jež jsou velmi silně korelované. Zdánlivě by neměl být velký problém rozdělení statistiky (5) odvodit. Bohužel toto rozdělení je natolik komplikované, že odpovídající kritické hodnoty (kvantily) mohou být spočteny pouze pro velmi malé hodnoty n.

V některých případech lze skutečné kritické hodnoty dostatečně přesně (z hlediska praktického použití) aproximovat. Jednou z nejjednodušších cest, jak je možno přibližné kritické hodnoty nalézt, je použití *Bonferroniho nerovnosti*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leqq \sum_{i=1}^{n} P(A_i), \tag{10}$$

kde $\{A_i\}$ jsou libovolné náhodné jevy. Odsud dostaneme

$$P\left(\max_{0\leqq k\leqq n-1} \left\{ \left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right| \right\} > C\right) \leqq \sum_{k=0}^{n-1} P\left\{ \left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right| > C \right\}.$$

Aplikujeme-li tento postup na testovou statistiku (5), potom nám $100\big(1-\alpha/(2n)\big)\%$ kvantil standardního normálního rozdělení $N(0,1)$ může posloužit jako horní odhad hledané kritické hodnoty na hladině významnosti α pro problém (4). Přibližné kritické hodnoty získané touto cestou jsou dobré pouze pro malé rozsahy výběru, tj. pro malé hodnoty n, ale jsou příliš konzervativní pro n velká. V takovém případě je lepší použít místo Bonferroniho nerovnosti výsledků o asymptotickém chování statistiky (5).

Užijeme-li zákon iterovaného logaritmu, lze dokázat, že

$$\max_{0\leqq k\leqq n-1} \left\{ \left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right| \right\} \longrightarrow \infty \quad \text{skoro jistě pro} \quad n \to \infty.$$

Z uvedeného vyplývá, že limitní rozdělení statistiky (5) neexistuje a že kritické hodnoty rostou do nekonečna, pokud $n \to \infty$. Problém je způsoben chováním posloupnosti $\left\{ (n-k)^{-1/2} \sum_{i=k+1}^{n} Y_i,\ k = 0, 1, \ldots, n-1 \right\}$ pro k blízká n. To byl také důvod, proč

někteří autoři navrhli používat místo statistik (5) a (6) testové statistiky tzv. *useknutého maximálního typu*:

$$\max_{0\leqq k\leqq \lfloor(1-\beta)n\rfloor}\left\{\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right\}, \qquad \max_{0\leqq k\leqq \lfloor(1-\beta)n\rfloor}\left\{\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right|\right\}, \tag{11}$$

kde β je malá kladná konstanta (menší než jedna) a $\lfloor x\rfloor$ označuje celou část x. Statistiky (11) lze použít v situaci, kdy je známo, že ke změně zcela jistě nedošlo v posledních $100\beta\,\%$ časových obdobích. Je výhodné, že statistiky (11) mají limitní rozdělení. Díky tomu můžeme k rozhodnutí, zda nulovou hypotézu zamítnout či nikoliv, použít kritickou hodnotu odpovídajícího limitního rozdělení.

Někteří autoři dále doporučují pracovat s maximem vážených statistik (9) nebo s jejich absolutními hodnotami, tj. se statistikami tvaru:

$$\max_{0\leqq k\leqq n-1}\left\{w(k)\cdot\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right\}, \qquad \max_{0\leqq k\leqq n-1}\left\{w(k)\cdot\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right|\right\}.$$

Váhy $w(k)$ odpovídají apriorní informaci o možném bodu změny, kterou známe ze zkušenosti s obdobným typem problémů nebo kterou nám poskytl nezávislý expert.

Přejděme nyní k troše teorie. Asymptotické chování maxima posloupnosti náhodných veličin je v případě námi uvažovaném dáno limitním gaussovským procesem. Tyto limitní procesy se naneštěstí pro různé testové statistiky liší, a tudíž se liší i pravděpodobnosti překročení úrovně (exceedence level probabilities). V uvedeném příkladu 1 je limitním procesem $\left\{\frac{W(1-t)}{\sqrt{1-t}},\, t\in[0,1)\right\}$, kde $\{W(t),\, t\geqq 0\}$ označuje Wienerův proces. Více se lze o Wienerově procesu, jeho vlastnostech a použití dozvědět z knihy Štěpán [27].

Přibližné kritické hodnoty pro statistiky (5) a (6) se mohou počítat pomocí pravděpodobnosti přechodu pro limitní proces takto:

$$P\left(\max_{0\leqq k\leqq n-1}\left\{\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right\} > \frac{x+b_n}{a_n}\right) \approx 1-\exp\Big\{-\frac{1}{2}\,\mathrm{e}^{-x}\Big\}, \quad x\in\mathbb{R}^1, \tag{12}$$

$$P\left(\max_{0\leqq k\leqq n-1}\left\{\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right|\right\} > \frac{x+b_n}{a_n}\right) \approx 1-\exp\left\{-\mathrm{e}^{-x}\right\}, \quad x\in\mathbb{R}^1, \tag{13}$$

kde

$$a_n=\sqrt{2\log\log n} \quad \text{a} \quad b_n=2\log\log n+\frac{1}{2}\log\log\log n-\frac{1}{2}\log\pi.$$

Přibližné kritické hodnoty pro useknuté statistiky maximálního typu (11) se mohou počítat z aproximací:

$$P\left(\max_{0\leqq k\leqq \lfloor(1-\beta)n\rfloor}\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i > x\right) \approx \frac{1}{2}\,x\varphi(x)\log\frac{1}{\beta}, \quad x\in\mathbb{R}^1, \tag{14}$$

$$P\left(\max_{0\leqq k\leqq \lfloor(1-\beta)n\rfloor}\left\{\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n} Y_i\right|\right\} > x\right) \approx x\varphi(x)\log\frac{1}{\beta}, \quad x\in\mathbb{R}^1. \tag{15}$$

Kritické hodnoty získané z (12) a (13) nejsou příliš dobré, neboť jsou příliš konzervativní. Naproti tomu aproximace (14) a (15) dávají mnohem přesnější přiblížení správných kritických hodnot. Bohužel tyto kritické hodnoty jsou významně ovlivněny volbou hodnoty β. Kritické hodnoty získané pomocí (14) a (15) lépe odpovídají skutečným kritickým hodnotám v případě, že můžeme useknout více časových okamžiků, tj. β je větší, což znamená, že máme přesnější představu o tom, kde ke změně mohlo dojít.

Pro statistiky součtového typu, viz například (7), jež získáme při použití pseudo-bayesovského přístupu, lze limitní kritické hodnoty získat díky konvergenci v distribuci (definici lze nalézt například v knize Štěpán [27])

$$\sum_{k=1}^{n} \frac{1}{n} \left(\frac{1}{\sqrt{n}} \sum_{i=k+1}^{n} Y_i \right)^2 \xrightarrow{\mathcal{D}} \int_0^1 W^2(t)\,\mathrm{d}t. \tag{16}$$

Rozdělení $\int_0^1 W^2(t)\,\mathrm{d}t$ studoval MacNeill [24, 25]. Poznamenejme, že aproximace (12)–(16) lze použít i pro nezávislá pozorování, která nemají normální rozdělení.

Výpočet kritických hodnot pro statistiku (8) je velmi jednoduchý, neboť tyto statistiky mají normální rozdělení $N\big(0,\, 1/3 - 1/(2n) + 1/(6n^2)\big)$.

Poznámka: Potřebné kritické hodnoty je též možno získat pomocí simulací. V práci Antoch et al. [5] lze nalézt mnoho simulovaných kritických hodnot pro metody již popsané spolu s mnohem podrobnějším popisem těchto metod.

Poté, co jsme si alespoň stručně vysvětlili některé základní myšlenky, se vraťme zpět k našim příkladům.

Příklad 1 (pokračování).

Jestliže náhodné chyby $\{e_i\}$ jsou nezávislé s týmž normálním rozdělením $N(0, \sigma^2)$, kde σ^2 neznáme, potom mají statistiky maximálního typu odvozené metodou maximální věrohodnosti tvar

$$\max_{1 \leqq k \leqq n-1} \left\{ \frac{1}{s_k} \left| \sqrt{\frac{n}{k(n-k)}} \sum_{i=1}^{k} (Y_i - \overline{Y}_n) \right| \right\} = \\ = \max_{1 \leqq k \leqq n-1} \left\{ \frac{1}{s_k} \left| \sqrt{\frac{k(n-k)}{n}} (\overline{Y}_k - \overline{Y}_k{}^o) \right| \right\}, \tag{17}$$

resp.

$$\max_{\lfloor \beta n \rfloor \leqq k \leqq \lfloor (1-\beta) n \rfloor} \left\{ \frac{1}{s_k} \left| \sqrt{\frac{n}{k(n-k)}} \sum_{i=1}^{k} (Y_i - \overline{Y}_n) \right| \right\} = \\ = \max_{\lfloor \beta n \rfloor \leqq k \leqq \lfloor (1-\beta) n \rfloor} \left\{ \frac{1}{s_k} \left| \sqrt{\frac{k(n-k)}{n}} (\overline{Y}_k - \overline{Y}_k{}^o) \right| \right\}, \tag{18}$$

kde

$$\overline{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i, \qquad \overline{Y}_k^{\,o} = \frac{1}{n-k}\sum_{i=k+1}^{n} Y_i,$$

$$s_k^2 = \frac{1}{n-2}\left(\sum_{i=1}^{k}\big(Y_i - \overline{Y}_k\big)^2 + \sum_{i=k+1}^{n}\big(Y_i - \overline{Y}_k^{\,o}\big)^2\right),$$

zatímco statistika součtového typu má tvar

$$\frac{1}{\widehat{\sigma}^{\,2}}\sum_{k=1}^{n}\frac{1}{n}\left(\frac{1}{\sqrt{n}}\sum_{i=k+1}^{n}\big(Y_i - \overline{Y}_n\big)\right)^2, \quad \text{kde} \quad \widehat{\sigma}^{\,2} = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \overline{Y}_n\big)^2.$$

Podobně jako v situaci, kdy byl rozptyl σ^2 znám, můžeme i zde nalézt přibližné kritické hodnoty použitím Bonferroniho nerovnosti (10). Pro pevné známé k je statistika

$$T_k = \frac{1}{s_k}\sqrt{\frac{k(n-k)}{n}}\,\big(\overline{Y}_k - \overline{Y}_k^{\,o}\big)$$

běžně užívanou testovou statistikou pro dvouvýběrový problém. Stačí si tedy pouze uvědomit, že všechny statistiky $\{T_k\}$ jsou rozděleny podle Studentova t-rozdělení s $n-2$ stupni volnosti.

Pro velká n mohou být asymptotické kritické hodnoty nalezeny aproximacemi ($x \in \mathbb{R}^1$, $a_n = \sqrt{2\log\log n}$, $b_n = 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log\pi$):

$$P\left(\max_{1\leqq k\leqq n-1}\left\{\frac{1}{s_k}\left|\sqrt{\frac{n}{k(n-k)}}\sum_{i=1}^{k}\big(Y_i - \overline{Y}_n\big)\right|\right\} > \frac{x+b_n}{a_n}\right) \approx 1 - \exp\big\{-2\mathrm{e}^{-x}\big\}, \quad (20)$$

$$P\left(\max_{\lfloor\beta n\rfloor\leqq k\leqq\lfloor(1-\beta)n\rfloor}\left\{\frac{1}{s_k}\left|\sqrt{\frac{n}{k(n-k)}}\sum_{i=1}^{k}\big(Y_i - \overline{Y}_n\big)\right|\right\} > x\right) \approx 2x\varphi(x)\log\frac{1-\beta}{\beta}\,. \quad (21)$$

Přibližné kritické hodnoty pro statistiky součtového typu mohou být získány z aproximací ($x \in \mathbb{R}^1$)

$$P\left(\frac{1}{\widehat{\sigma}^{\,2}}\sum_{k=1}^{n}\frac{1}{n}\left(\frac{1}{\sqrt{n}}\sum_{i=k+1}^{n}\big(Y_i - \overline{Y}_n\big)\right)^2 > x\right) \approx$$

$$\approx 1 - \frac{\sqrt{2}}{\pi^{3/2}\sqrt{x}} \times \sum_{j=0}^{\infty}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\sqrt{2j+\frac{1}{2}}\,\exp\left\{-\frac{(4j+1)^2}{16x}\right\}K_{1/4}\left(\frac{(4j+1)^2}{16x}\right),$$

kde $K_{1/4}(\cdot)$ označuje modifikovanou Besselovu funkci druhého typu (viz například funkce `besselk(nu,z)` v Matlabu nebo `BesselK` v Mathematice).

V případě dat z Gütsche a Säntisu statistika (17) z příkladu 1, resp. statistika (18), kde $\beta = 0{,}05$, nabývá hodnoty 12,9, jež je vysoce významná bez ohledu na to, jakou aproximaci kritické hodnoty použijeme. Znamená to, že sledované rozdíly $\{Y_i\}$ nekolísají kolem stále stejné hodnoty.

Příklad 2 (pokračování).

Statistiky maximálního typu pro problém (2) mají tvar

$$\max_{1\leqq k<n}\left\{\frac{\widehat{b}_k}{s(\widehat{b}_k)}\right\} \quad \text{a} \quad \max_{1\leqq k\leqq\lfloor(1-\beta)n\rfloor}\left\{\frac{\widehat{b}_k}{s(\widehat{b}_k)}\right\}, \tag{22}$$

kde $\widehat{b}_k$ je odhad b získaný metodou nejmenších čtverců a $s(\widehat{b}_k)$ označuje odhad standardní odchylky $\widehat{b}_k$ za předpokladu, že se změna vyskytla v čase k. Testovou statistiku $\widehat{b}_k/s\big(\widehat{b}_k\big)$ bychom mohli použít při testování nulové hypotézy H_2 proti alternativě A_2 v případě, že bychom čas změny m předem znali a věděli, že $\mathsf{m}=k$. Jde o jednoduchý problém lineární regrese, popsaný například v knize Anděl [3]. Pro aplikace Bonferroniho nerovnosti (10) si všimněme, že pokud chyby jsou nezávislé, stejně rozdělené s normálním rozdělením, potom za platnosti H_2 se veličiny

$$\frac{\widehat{b}_k}{s(\widehat{b}_k)}=\frac{\dfrac{1}{s_k}\dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=k+1}^{n}\big(Y_i-\overline{Y}_n\big)\frac{i-k}{n}}{\sqrt{\dfrac{(n-k)(n-k+1)(n-k+\frac{1}{2})}{3n^3}-\dfrac{(n-k)^2(n-k+1)^2}{4n^4}}}, \quad k=1,\ldots,n-1,$$

řídí t-rozdělením s $n-2$ stupni volnosti, přičemž s_k, $k=1,\ldots,n-1$, jsou obvyklé odhady σ založené na reziduálním součtu čtverců.

Pro velká n lze kritické hodnoty získat aproximacemi ($x\in\mathbb{R}^1$):

$$P\left(\max_{1\leqq k\leqq n-1}\left\{\frac{\widehat{b}_k}{s(\widehat{b}_k)}\right\}>\frac{x+b_{n,2}}{a_n}\right)\approx 1-\exp\big\{-\mathrm{e}^{-x}\big\}, \tag{23}$$

kde

$$a_n=\sqrt{2\log\log n} \quad \text{a} \quad b_{n,2}=2\log\log n+\log\frac{\sqrt{3}}{4\pi},$$

resp.

$$P\left(\max_{1\leqq k\leqq\lfloor(1-\beta)n\rfloor}\left\{\frac{\widehat{b}_k}{s(\widehat{b}_k)}\right\}>x\right)\approx\frac{1}{\sqrt{\pi}}\,2\varphi(x)\int_0^{1-\beta}\frac{\sqrt{6t}}{(1-t)(1+3t)}\,\mathrm{d}t.$$

Obě statistiky (22) pro Jonesovu teplotní řadu nabývají hodnoty 17,7, jež je opět vysoce významná bez ohledu na použitou aproximaci kritických hodnot.

Příklad 3 (pokračování).

Statistiky maximálního typu pro problém (3) jsou tvaru

$$\max_{2\leqq k\leqq n-2}\big\{F_k\big\} \quad \text{a} \quad \max_{\lfloor\beta n\rfloor\leqq k\leqq\lfloor(1-\beta)n\rfloor}\big\{F_k\big\}, \tag{24}$$

kde

$$F_k=\frac{1}{s_k^2}\left(\frac{nk\big(\overline{Y}_k-\overline{Y}_n\big)^2}{n-k}+\frac{Q_{xy}^2(k)}{Q_{xx}(k)}+\frac{Q_{xy}^{o2}(k)}{Q_{xx}^o(k)}-\frac{Q_{xy}^2(n)}{Q_{xx}(n)}\right);$$

s_k^2 jsou obvyklé odhady rozptylu náhodných chyb σ^2 založené na reziduálním součtu čtverců za předpokladu, že změna se vyskytla v čase k, a

$$Q_{xx}(k) = \sum_{i=1}^{k} \big(x_i - \overline{x}_k\big)\big(x_i - \overline{x}_k\big), \qquad Q_{xy}(k) = \sum_{i=1}^{k} \big(x_i - \overline{x}_k\big)\big(Y_i - \overline{Y}_k\big),$$
$$Q_{xx}^o(k) = \sum_{i=k+1}^{n} \big(x_i - \overline{x}_k^{\,o}\big)\big(x_i - \overline{x}_k^{\,o}\big), \qquad Q_{xy}^o(k) = \sum_{i=k+1}^{n} \big(x_i - \overline{x}_k^{\,o}\big)\big(Y_i - \overline{Y}_k^{\,o}\big).$$

Jestliže náhodné chyby $\{e_i\}$ jsou nezávislé, stejně rozdělené s normálním rozdělením, potom za platnosti H_3 se všechny veličiny $\{F_k : k = 2, \dots, n-2\}$ řídí F-rozdělením s 2 a $n-4$ stupni volnosti. Pro velké hodnoty n mohou být kritické hodnoty spočteny aproximací ($x \in \mathbb{R}^1$):

$$P\left(\max_{2 \leqq k \leqq n-2} \big\{F_k\big\} > \left(\frac{x + b_{n,3}}{a_n}\right)^2\right) \approx 1 - \exp\big\{-2\mathrm{e}^{-x}\big\}, \tag{25}$$

kde

$$a_n = \sqrt{2 \log\log n} \quad \text{a} \quad b_{n,3} = 2\log\log n + \frac{1}{2}\log\log\log n,$$

resp.

$$P\left(\max_{\lfloor \beta n \rfloor \leqq k \leqq \lfloor (1-\beta) n \rfloor} \big\{F_k\big\} > x\right) \approx x\mathrm{e}^{-\frac{1}{2}x} \log\frac{1-\beta}{\beta} \quad (x > 0). \tag{26}$$

Statistika (24) nabývá pro data z příkladu 3 hodnoty 31,78. Použijeme-li Bonferroniho nerovnost, potom horní odhad tzv. p-hodnoty je roven $8{,}24 \cdot 10^{-7}$, takže můžeme zamítnout nulovou hypotézu, že vztah mezi srážkami a odtoky je stacionární.

Vedle metody maximální věrohodnosti je pro rozhodnutí, zda existuje lineární vztah mezi závisle a nezávisle proměnnou, též velmi populární tzv. *metoda rekurzivních reziduí*, kterou navrhli Brown et al. [9]. Metoda je založena na rozdílech mezi předpovědí hodnoty časové řady o jeden krok dopředu a skutečně naměřenou hodnotou. Pokud součet takových odchylek (rekurzivních reziduí) je příliš velký, pak to ukazuje na nestacionaritu sledované řady. Velká přednost této metody spočívá v tom, že není úzce vázána na určité specifické porušení stacionarity, jako je tomu u popsaných metod. Vysvětlení této metody však přesahuje rámec tohoto článku.

3. Odhady bodu změny

Nejpopulárnější metodou pro odhadování bodu změny je metoda maximální věrohodnosti, popsaná například v knize Anděl [3]. Jestliže došlo ke změně ve střední hodnotě normálního rozdělení, jako je tomu v našich příkladech, potom jde o odhad parametrů v nelineárním regresním modelu, pro jehož nalezení lze užít metodu nejmenších čtverců. V případě, kdy se studovaná posloupnost neřídí normálním rozdělením, je lepší užít některý neparametrický postup nebo robustní odhad. Pokud máme k dispozici apriorní rozdělení bodu změny, je vhodné použít bayesovský přístup.

Statistici obvykle nejsou spokojeni s pouhým bodovým odhadem bodu změny, nýbrž dávají přednost intervalu spolehlivosti. Je třeba říci, že přesné rozdělení odhadu bodu změny je téměř vždy natolik složité, že ke konstrukci intervalu spolehlivosti je třeba užít rozdělení asymptotické. Bohužel limitní rozdělení se podstatně liší model od modelu a rovněž závisí na tom, zda jde o změnu náhlou nebo postupnou atd.

Příklad 1 (pokračování).

Označme $\widehat{\mu}$, $\widehat{\mathsf{m}}$ a $\widehat{\delta}$ odhady metodou nejmenších čtverců parametrů μ, m a δ, tj. hodnoty, které minimalizují

$$\sum_{i=1}^{k}\big(Y_i-\mu\big)^2+\sum_{i=k+1}^{n}\big(Y_i-\mu-\delta\big)^2, \tag{27}$$

a $\widehat{\sigma}^{\,2}$ nechť je odhad σ^2, tj.

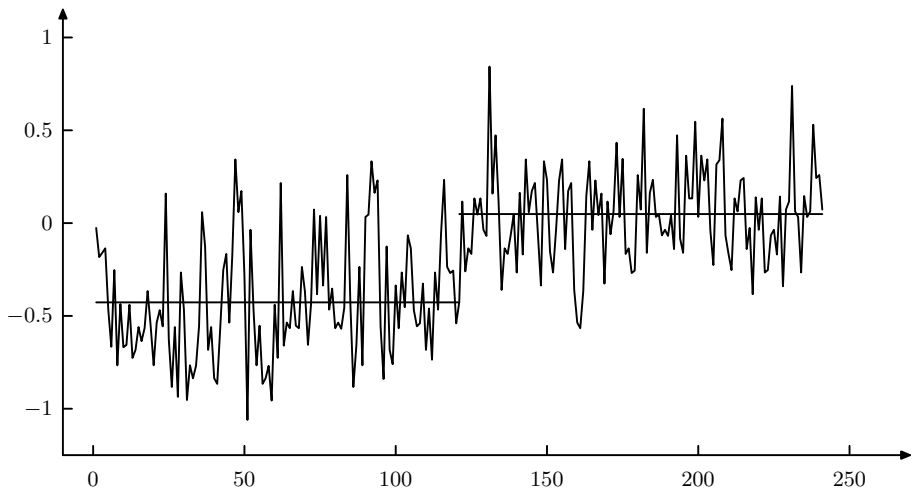
$$s^2_{\widehat{m}}=\frac{1}{n-2}\left\{\sum_{i=1}^{\widehat{\mathsf{m}}}\big(Y_i-\overline{Y}_{\widehat{\mathsf{m}}}\big)^2+\sum_{i=\widehat{\mathsf{m}}+1}^{n}\big(Y_i-\overline{Y}^{\,o}_{\widehat{\mathsf{m}}}\big)^2\right\}. \tag{28}$$

Potom pro malé hodnoty δ platí

$$P\left(\frac{\widehat{\delta}^{\,2}}{s^2_{\widehat{m}}}\,\big(\widehat{\mathsf{m}}-\mathsf{m}\big)\leqq x\right)\approx P(V\leqq x),\quad x\in\mathbb{R}^1, \tag{29}$$

kde

$$P\big(V\leqq x\big)=\begin{cases}1+\sqrt{\dfrac{x}{2\pi}}\,\mathrm{e}^{-\frac{1}{8}x}-\dfrac{x+5}{2}\,\Phi\Big(-\dfrac{1}{2}\sqrt{x}\Big)+\dfrac{3}{2}\,\mathrm{e}^{x}\Phi\Big(-\dfrac{3}{2}\,\sqrt{x}\Big), & x\geqq 0,\\ 1-P\big(V\leqq -x\big), & x<0.\end{cases} \tag{30}$$



Obr. 1. Data a model pro Gütsch-Säntis.

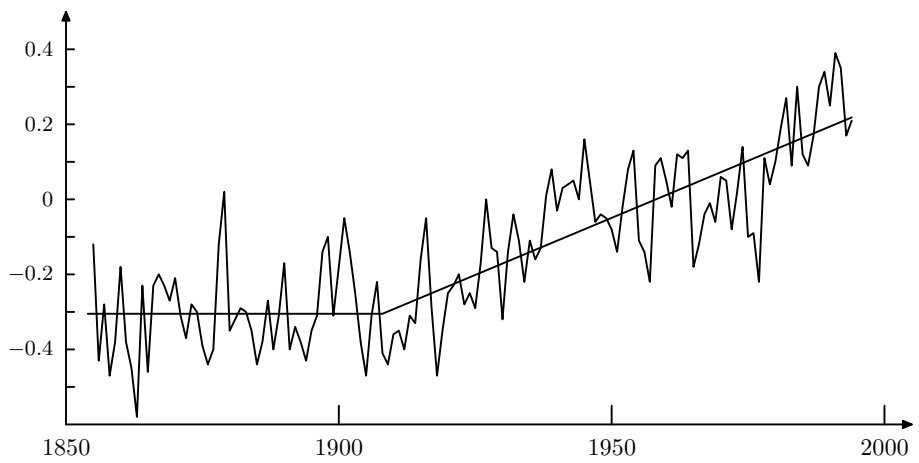
Pro data z Gütsche a Säntisu dostáváme následující odhady: $\widehat{\mathsf{m}} = 120$, $\widehat{\delta} = 0{,}4791$ a $\widehat{\sigma}^{\,2} = 0{,}083$. Aplikujeme-li (29) a vezmeme v úvahu, že 97,5% kvantil rozdělení (30) je 11,033, dostáváme, že se spolehlivostí 95 % se změna vyskytla mezi 116. a 124. pozorováním.

Příklad 2 (pokračování).

Označme $\widehat{\mathsf{m}}$ a $\widehat{b}_{\widehat{\mathsf{m}}}$ odhady metodou nejmenších čtverců pro m a b, a $s_{\widehat{\mathsf{m}}}$ odhad σ založený na reziduích. Potom platí

$$P\left(\frac{\widehat{b}_{\widehat{\mathsf{m}}}}{s\big(\widehat{b}_{\widehat{m}}\big)}\,\frac{\widehat{\mathsf{m}}-\mathsf{m}}{\sqrt{n}}\,\sqrt{\frac{(\widehat{\mathsf{m}}/n)(1-\widehat{\mathsf{m}}/n)}{1+3\widehat{\mathsf{m}}/n}} \leqq x\right) \approx \Phi(x) \qquad (x \in \mathbb{R}^1). \tag{31}$$

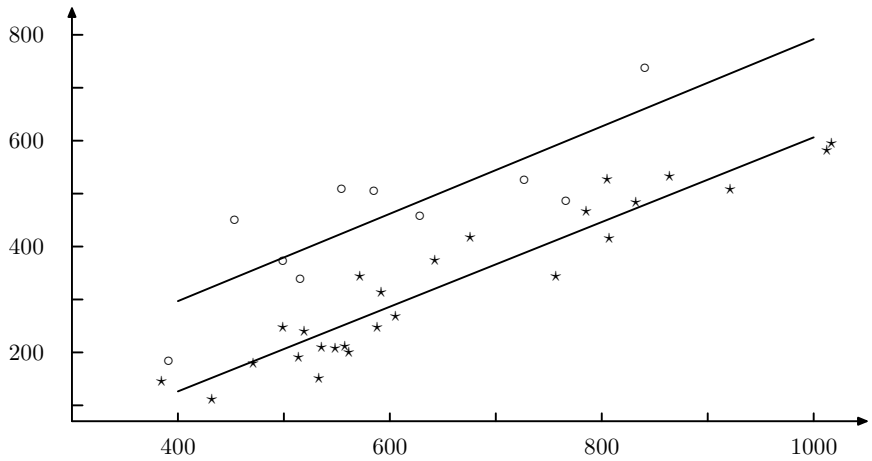
Použijeme-li (31) s $\widehat{\mathsf{m}} = 54$, což odpovídá roku 1907, $\widehat{b}_{\widehat{\mathsf{m}}} = 0{,}694$ a $s_{\widehat{\mathsf{m}}} = 0{,}117$, můžeme se spolehlivostí 95 % učinit závěr, že lineární trend se objevil mezi léty 1906–1908. Obrázek 2 ukazuje data spolu s „optimálním“ modelem.



Obr. 2. Data a model pro Jonesovu řadu.

Příklad 3 (pokračování).

Odhad bodu změny m získaný metodou nejmenších čtverců je roven $\widehat{\mathsf{m}} = 26$. Existují též asymptotické metody, které umožňují nalézt limitní rozdělení $\widehat{\mathsf{m}}$, blíže viz Antoch et al. [5]. Naneštěstí počet pozorování, která máme k dispozici ($n = 36$), je příliš malý pro jejich použití. Obrázek 3 ukazuje lineární vztah mezi srážkami a odtoky jak v prvním, tak v druhém časovém období, tj. $y = -193{,}6 + 0{,}8\,x$ a $y = -33{,}1 + 0{,}82\,x$. Symbolem „$\star$“ jsou označena pozorování vztahující se k prvním 26 letům pozorování, zatímco symbolem „$\circ$“ pozorování z posledních deseti let.



Obr. 3. Data a model pro Malou Ráztoku.

4. Závislost

Jedním z nejtypičtějších rysů hydrologických a meteorologických dat je závislost mezi časově blízkými pozorováními. Závislost má obecně velký vliv na rozhodnutí, zda řada je stacionární, a na délku intervalu spolehlivosti pro odhad bodu změny. V modelech odpovídajících příkladům 1 a 2 je možné za předpokladu, že studovaná řada tvoří ARMA posloupnost, ukázat, že kritické hodnoty pro testové statistiky odvozené principem maximální věrohodnosti musí být vynásobeny konstantou $\sqrt{2\pi h(0)/c}$, kde $h(\cdot)$ označuje spektrální hustotu procesu $\{e_t\}$ a $c = \mathsf{var}\ e_t$. Hranice konfidenčních intervalů je přitom třeba adaptovat odpovídajícím způsobem. Detailnější popis lze najít například v pracích Antoch et al. [6] nebo Bai [7].

5. Vícenásobné změny

Je-li počet možných změn známý, můžeme opět použít metodu maximální věrohodnosti. Naproti tomu je-li počet změn neznámý, je určení jejich počtu a míst velice obtížné. Nejčastěji se zde používá kombinace statistických postupů a informačních kritérií. Pokud si můžeme být (více či méně) jisti, že okamžiky změn se nevyskytují příliš blízko sebe, ukázaly se vhodným nástrojem modifikované neparametrické postupy využívající klouzavé okénko, jímž se na data díváme. Pro situace popsané v příkladu 1 lze často s úspěchem použít metodu sekvenčního dělení studované posloupnosti dat. Yao a Au [31] doporučují naopak určit i zde počet změn pomocí informačního kritéria.

Poděkování. Tento příspěvek byl připraven za podpory grantů GA ČR 201/00/0769, MSM 210000001, MSM 321100008. Autoři dále chtějí poděkovat kolegovi Janu Klaschkovi za velmi pečlivé přečtení tohoto článku a podnětné připomínky.

L i t e r a t u r a

[1] Albin, J. M. P.: *On extremal theory for stationary processes.* Ann. Probab. *18* (1990), 92–108.

[2] Alexandersson, H.: *A homogeneity test applied to precipitation data.* J. Climatol. *6* (1986), 661–675.

[3] Anděl, J.: *Matematická statistika.* SNTL/ALFA, Praha 1985.

[4] Antoch, J., Hušková, M., Jarušková, D.: *Change-point problém po deseti letech.* ROBUST'1998. Antoch, J. a Dohnal, G. eds., JČMF, Praha 1998, 1–42.

[5] Antoch, J., Hušková, M., Jarušková, D.: *Off-line process control.* Multivariate Total Quality Control: Foundation and Recent Advances. Springer-Verlag, Heidelberg 2002, 1–86.

[6] Antoch, J., Hušková, M., Prášková, Z.: *Effect of dependence on statistics for determination of change.* J. Stat. Plan. Infer. *60* (1997), 291–310.

[7] Bai, J.: *Least squares estimation of a shift in linear processes.* J. Time Series Analysis *15* (1994), 453–472.

[8] Bhattacharya, P. K.: *Weak convergence of the log-likelihood process in the two-phase regression problem.* Proc. of the R. C. Bose Symposium on Probability, Statistics and Design of Experiments, Wiley Eastern, New Delhi 1990, 145–156.

[9] Brown, R. L., Durbin, J., Evans, J. M.: *Techniques for testing the constancy of regression relationships over time (with discussion).* JRSS *B 37* (1975), 149–192.

[10] Buishand, T. A.: *Tests for detecting a shift in the mean of hydrological records.* J. Hydrol. *73* (1984), 51–69.

[11] Chernoff, H., Zacks, S.: *Estimating the current mean of normal distribution which is subjected to changes in time.* Ann. Math. Statist. *35* (1964), 999–1018.

[12] Csörgő, M., Horváth, L.: *Limit Theorems in Change Point Analysis.* J. Wiley, New York 1997.

[13] Deshayes, J., Picard, D.: *Off-line statistical analysis of change point models using nonparametric and likelihood methods.* Lecture Notes in Control and Information Sciences 77, Basseville, M. et al. eds., Springer Verlag, New York 1986, 103–168.

[14] Gardner, L. A.: *On detecting changes in the mean of normal variates.* Ann. Math. Statist. *40* (1969), 116–126.

[15] Hátle, J., Likeš, J.: *Základy počtu pravděpodobnosti a matematické statistiky.* SNTL/ALFA, Praha 1974.

[16] Hinkley, D. V.: *Inference about the intersection in two-phase regression.* Biometrika *56* (1969), 495–504.

[17] Chlebek, A. Jařabáč, M.: *Vliv pokračujících těžeb porostů v povodí a obnovy na odtok vody.* Zpráva pro závěrečné oponentní řízení, Výzkumný ústav lesního hospodářství a myslivosti, Jíloviště-Strnady 1989.

[18] James, B., James, K. L., Siegmund, D.: *Tests for change points.* Biometrika *74* (1987), 71–84.

[19] Jarušková, D.: *Change-point detection for dependent data and application to hydrology.* Istatistik. Journal of the Turkish Statistical Association *1* (1998), 9–21.

[20] Jarušková, D.: *Change point detection in meteorological measurements.* Monthly Weather Review *124* (1996), 1535–1543.

[21] Jarušková, D.: *Some problems with application of change point detection methods to environmental data.* Environmetrics *8* (1997), 469–483.

[22] Jones, P. D., Wigley, M. L., Briffa, K. R.: *Global and hemispheric temperature anomalies — land and maritime instrumental records.* Boden, T. A., Kaiser, D. P., Sepanski, R. J., Stoss, F. W., eds. Trends '93: A Compendium of Data on Global Change, ORNC/CDIAC-65, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tenn., USA 1994, 603–608.

[23] Kander, Z., Zacks, S.: *Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points.* Ann. Math. Statist. *37* (1966), 1196 až 1210.
[24] MacNeill, I. B.: *Tests for change of parameter at unknown time and distribution of some related functionals of Brownian motion.* Ann. Statist. *2* (1974), 950–962.
[25] MacNeill, I. B.: *Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times.* Ann. Statist. *6* (1978), 422–433.
[26] Rhoadas, D. A., Salinger, M. J.: *Adjustment of temperature and rainfalls records for site changes.* J. Climatol. *13* (1993), 899–913.
[27] Štěpán, J.: *Teorie pravděpodobnosti. Matematické základy.* Academia, Praha 1987.
[28] Vannitsem, S., Nicolis, C.: *Detecting climatic transitions: Statistical and dynamical aspects.* Beitr. Phys. Atmosph. *64* (1991), 245–254.
[29] Worsley, K. J.: *Testing for a two-phase multiple regression.* Technometrics *25* (1983), 35–42.
[30] Yao Yi-Ching, Davis, R. A.: *The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates.* Sankhya *48* (1986), 339–353.
[31] Yao Yi-Ching, Au, S. T.: *Least-squares estimation of a step function.* Sankhya *51* (1989), 370–381.

Informace a entropie z pohledu fyzika

Milan Marvan, Praha

1. Úvod

Entropie jako fyzikální pojem byla zavedena již v 19. století nejdříve ve fenomenologické termodynamice při studiu účinnosti tepelných strojů. O statisticko-fyzikální interpretaci pojmu entropie (odst. 2) se ještě v témž století zasloužil L. Boltzmann, který ukázal souvislost pojmu entropie s pravděpodobností. Spíše než jeho matematický vzorec pronikla do širšího povědomí jeho názorná „definice“ entropie. Podle této definice je entropie mírou neuspořádanosti stavu systému. Poznamenejme, že pro entropii existuje i druhá názorná interpretace: entropie je míra neurčitosti podrobného stavu (mikrostavu) systému. Poslední interpretace je bližší pojetí, které je známo v teorii informace, nikoliv ve fyzice.

Doc. RNDr. Milan Marvan, CSc. (1932), katedra makromolekulární fyziky, MFF UK, V Holešovičkách 2, 180 00 Praha 8, e-mail:marvan@kmf.troja.mff.cuni.cz