Matematicko-fyzikální fakulta
Univerzita Karlova

Martin Strnad

Sekvenční monitorování změny v lineárním modelu

Statistický seminář 1

3. května 2023

Motivace

- Situace: pozorujeme data s určitou frekvencí (z oblasti ekonomie, financí, geofyziky apod.)
- Cíl: od nějakého okamžiku m začneme při sběru testovat, zda-li data stále odpovídají modelu před časem m
- Kompromis mezi zpožděním detekce a falešným poplachem
- $Y_i = X_i^T \beta_i + \varepsilon_i, \ i \in \mathbb{N}$
- $X_i = (1, X_{i2}, \ldots, X_{ip})^T, \ i \in \mathbb{N}$
- $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2), \ i \in \mathbb{N}$

Hypotéza

- Žádná změna v β v historických datech: $\beta_i = \beta_0,\ i = 1, \ldots, m$
- Testujeme

$$\mathrm{H}_0 : \beta_i = \beta_0,\ i = m+1, m+2, \ldots$$

proti

$$\mathrm{H}_1 : \exists k^* \in \mathbb{N}\ \exists \beta_* \neq \beta_0 :\ \beta_i = \begin{cases} \beta_0, & m < i < m + k^*, \\ \beta_*, & i \geq m + k^*. \end{cases}$$

Způsob testování

- Idea: najít vhodnou statistiku (detektor) $\Gamma(m,k)$ a vhodnou hraniční (rozhodovací) funkci $g(m,k)$
- Zamítáme H_0 v čase $\mathrm{stop}(m) = \inf\{k \in \mathbb{N} \,\big|\, \Gamma(m,k) \geq g(m,k)\}$
- Je třeba najít Γ a g tak, aby pro předepsané α platilo

$$\lim_{m\to\infty} \mathrm{P}[\mathrm{stop}(m) < \infty \mid \mathrm{H}_0] = \alpha, \quad \lim_{m\to\infty} \mathrm{P}[\mathrm{stop}(m) < \infty \mid \mathrm{H}_1] = 1.$$

- Co vůbec zde znamená $m \to \infty$?

Testové statistiky

- $\widehat{\beta}_n := \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \sum_{i=1}^n X_i Y_i = \left(\mathbb{X}_n^T \mathbb{X}_n\right)^{-1} \mathbb{X}_n^T \mathbb{Y}_n$
- $\sigma_m^2 := (m-p)^{-1} ||\mathbb{Y}_m - \mathbb{X}_m \widehat{\beta}_m||^2$
- $\widehat{\varepsilon}_i := Y_i - X_i^T \widehat{\beta}_m,\ i = 1, 2, \ldots, m, m+1, \ldots$
- $\widetilde{\varepsilon}_i := Y_i - X_i^T \widehat{\beta}_{i-1},\ i = 2, \ldots, m, m+1, \ldots$

$$\widehat{Q}(m,k) := \sum_{i=m+1}^{m+k} (Y_i - X_i^T \widehat{\beta}_m), \quad \widetilde{Q}(m,k) := \sum_{i=m+1}^{m+k} (Y_i - X_i^T \widehat{\beta}_{i-1})$$

Další předpoklady pro asymptotiku

- Dodatečné předpoklady:

$$\mathbb{E}[|\varepsilon_1|^{2+\nu}] < \infty \quad \text{pro nějaké } \nu > 0$$

$$\{\varepsilon_i \mid i \in \mathbb{N}\} \perp \{X_i \mid i \in \mathbb{N}\}$$

Existuje $\tau > 0$ a $C \rhd 0$ takové, že

$$\left\| \frac{1}{n} \mathbb{X}_n^T \mathbb{X}_n - C \right\| = \mathrm{O}(1/n^\tau) \quad \text{s.j.}$$

Asymptotika pro $\sum \widehat{\varepsilon}_i$ za H_0

Věta 1.1

Za zmíněných předpokladů a platnosti nulové hypotézy máme pro parametr $0 \leq \gamma < \min(\tau, 1/2)$

$$\mathrm{P}\left[\frac{1}{\widehat{\sigma}_m} \sup_{k \in \mathbb{N}} \frac{\left|\sum_{i=m+1}^{m+k} \widehat{\varepsilon}_i\right|}{\sqrt{m}\left(1+\frac{k}{m}\right)\left(\frac{k}{k+m}\right)^{\gamma}} \leq c\right] \overset{m\to\infty}{\longrightarrow} \mathrm{P}\left[\sup_{[0,1]} \frac{|W(t)|}{t^{\gamma}} \leq c\right].$$

Simulace - kritické hodnoty

Critical values $c_\alpha(\gamma)$ based on 50,000 replications of X_γ, (4.1)

γ	α				
	0.010	0.025	0.050	0.100	0.250
0.00	2.7912	2.4948	2.2365	1.9497	1.5213
0.15	2.8516	2.5475	2.2996	2.0273	1.6126
0.25	2.9445	2.6396	2.3860	2.1060	1.7039
0.35	3.0475	2.7394	2.5050	2.2433	1.8467
0.45	3.3015	3.0144	2.7992	2.5437	2.1729
0.49	3.5705	3.2944	3.0722	2.8259	2.4487

The Wiener process was approximated on a grid of 10,000 equi-spaced points in [0, 1].

Recall that $\{W(t),\ 0 \leqslant t < \infty\}$ is a Wiener process and define

$$X_\gamma = \sup_{0 \leqslant t \leqslant 1} |W(t)|/t^\gamma.$$

Věta 1.2

Za zmíněných předpokladů, $0 \leq \gamma < \min(\tau, 1/2)$ a navíc

$$c_1^T(\beta_0 - \beta_*) \neq 0,$$

máme za platnosti alternativní hypotézy

$$\frac{1}{\widehat{\sigma}_m} \sup_{k \in \mathbb{N}} \frac{\left| \sum_{i=m+1}^{m+k} \widehat{\varepsilon}_i \right|}{\sqrt{m}\left(1 + \frac{k}{m}\right)\left(\frac{k}{k+m}\right)^{\gamma}} \xrightarrow[m \to \infty]{\mathrm{P}} \infty.$$

Asymptotika pro $\sum \widetilde{\varepsilon}_i$ za H_0

Věta 2.1

Za zmíněných předpokladů, $0 \leq \gamma < \min(\tau, 1/2)$, máme za platnosti nulové hypotézy

$$\mathrm{P}\left[\frac{1}{\widehat{\sigma}_m} \sup_{k\in\mathbb{N}} \frac{\left|\sum_{i=m+1}^{m+k} \widetilde{\varepsilon}_i\right|}{\sqrt{m}h\left(\frac{k}{m}\right)} \leq 1\right] \overset{m\to\infty}{\longrightarrow} \mathrm{P}\left[\sup_{(0,\infty)} \frac{|W(t)|}{h(t)} \leq 1\right],$$

kde h je dostatečně „regulární“ kladná funkce.

Věta 2.2

Předpokladájme, že platí všechny příslušné předpoklady a navíc existuje $\widetilde{k} \in \mathbb{N}$ takové, že

$$\frac{m + k^*}{\sqrt{m} h(\widetilde{k}/m)} \left| \sum_{i=m+k^*}^{m+\widetilde{k}} \frac{X_i^T(\beta_0 - \beta_*)}{i} \right| \to \infty.$$

Potom za alternativy máme

$$\frac{1}{\widehat{\sigma}_m} \sup_{k \in \mathbb{N}} \frac{\left| \sum_{i=m+1}^{m+k} \widehat{\varepsilon}_i \right|}{\sqrt{m} h(k/m)} \xrightarrow[m \to \infty]{\mathrm{P}} \infty.$$

Simulace - chyba prvního druhu

Empirical sizes of tests based on the detectors D_γ, (4.2), $\gamma = 0, 0.25, 0.45$ and $\tilde{D}_a$, (4.3)

	q	$\gamma = 0.00$		$\gamma = 0.25$		$\gamma = 0.45$		Recursive (%)	
		10%	5%	10%	5%	10%	5%	10%	5%
$m = 25$	$2m$	1.92	0.72	4.56	1.76	5.76	3.12	0.56	0.24
	$4m$	6.00	2.88	8.72	4.24	7.68	4.16	2.60	1.32
	$6m$	7.52	3.84	9.64	4.92	8.28	4.52	3.84	2.24
	$9m$	8.56	4.52	10.56	5.28	8.64	4.72	5.40	3.12
$m = 100$	$2m$	1.28	0.32	3.72	1.52	5.80	3.16	0.24	0.08
	$4m$	5.88	2.20	6.80	3.12	7.04	3.84	2.00	0.80
	$6m$	7.84	3.60	7.80	3.92	7.44	3.92	3.04	1.20
	$9m$	9.20	4.56	8.68	4.52	7.60	4.04	4.20	1.68
$m = 300$	$2m$	1.12	0.24	4.20	1.52	5.96	2.36	0.64	0.12
	$4m$	4.56	1.84	7.48	3.72	7.52	2.92	2.24	0.72
	$6m$	6.48	2.72	8.36	4.28	7.84	3.16	3.32	1.40
	$9m$	7.84	3.48	9.04	4.56	8.20	3.28	4.52	1.92

For each m, the percentage of rejections for the process observed until time q is reported. The results are based on 2500 replications.

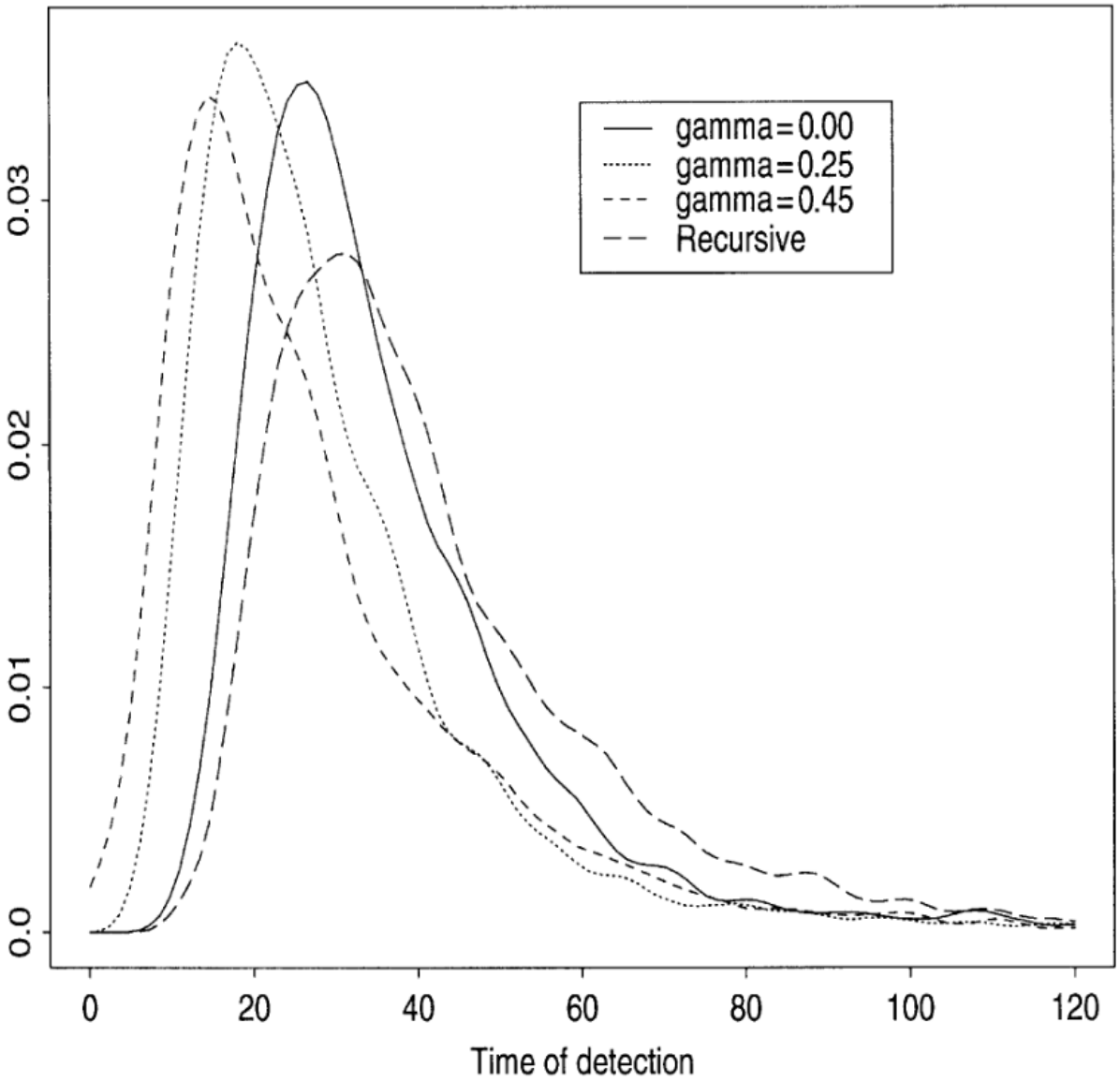
Simulace - distribuce času změny $\mu : 0 \rightarrow 0.8$

Five number summary for the distribution of the detection time for monitoring schemes with $\alpha = 0.10$

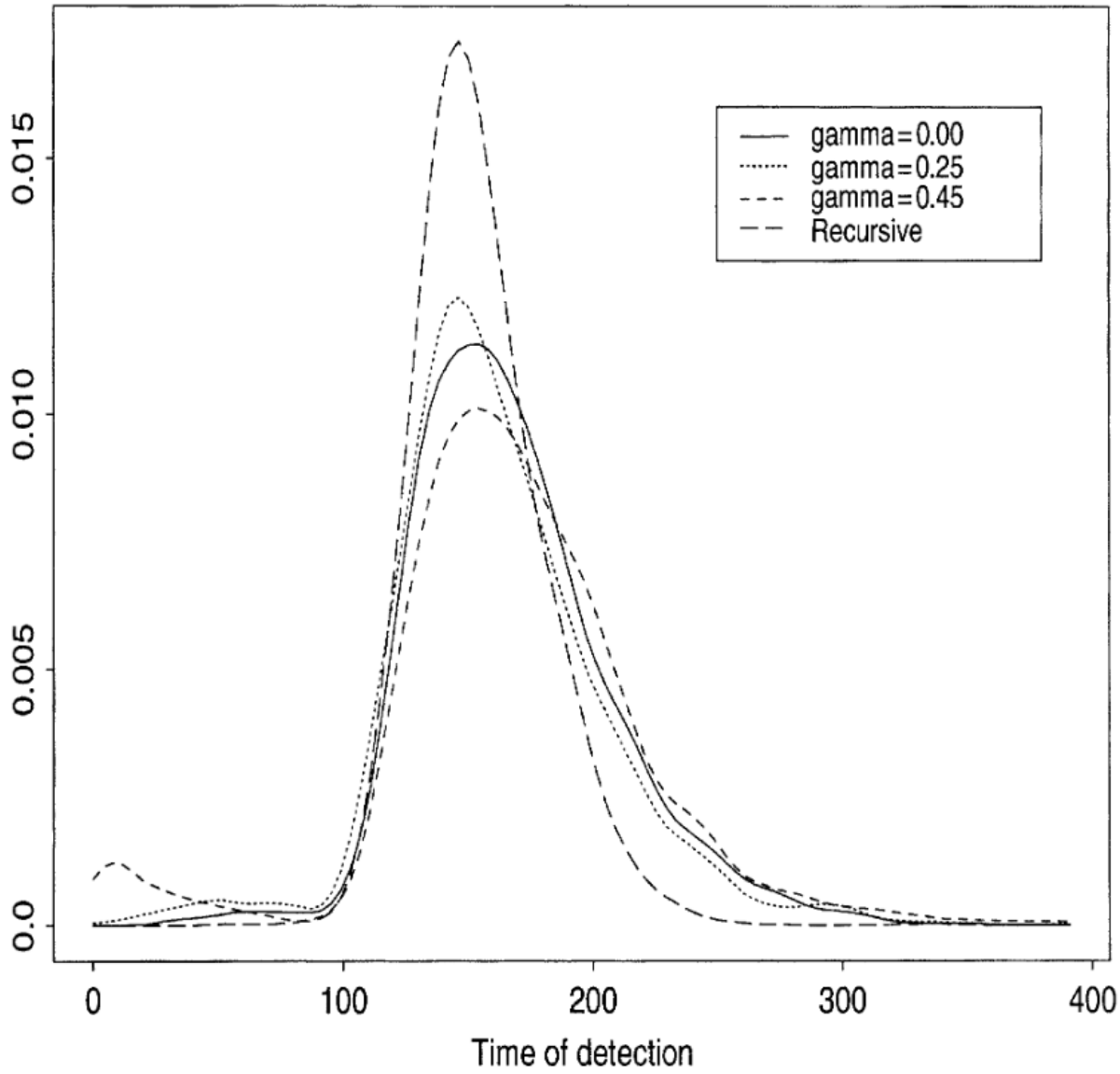
	$m = 25,\ k^* = 1$					$m = 100,\ k^* = 1$			
	$D_{0.00}$	$D_{0.25}$	$D_{0.45}$	$\tilde{D}_a$		$D_{0.00}$	$D_{0.25}$	$D_{0.45}$	$\tilde{D}_a$
min	4	1	1	5	min	12	3	1	12
Q1	15	19	7	19	Q1	26	15	8	31
med	22	16	15	35	med	33	20	13	39
Q3	37	33	35	65	Q3	41	27	20	49
max	200	200	200	200	max	123	98	136	145
	$m = 50,\ k^* = 5$					$m = 100,\ k^* = 100$			
	$D_{0.00}$	$D_{0.25}$	$D_{0.45}$	$\tilde{D}_a$		$D_{0.00}$	$D_{0.25}$	$D_{0.45}$	$\tilde{D}_a$
min	10	6	1	11	min	31	8	1	57
Q1	24	17	14	28	Q1	140	136	140	138
med	31	25	22	37	med	162	156	165	151
Q3	43	35	36	52	Q3	189	183	195	168
max	400	400	400	400	max	350	394	489	336

The results are based on 2500 replications.

Simulace - distribuce času změny $\mu : 0 \rightarrow 0.8$



Simulace - distribuce času změny $\mu : 0 \rightarrow 0.8$

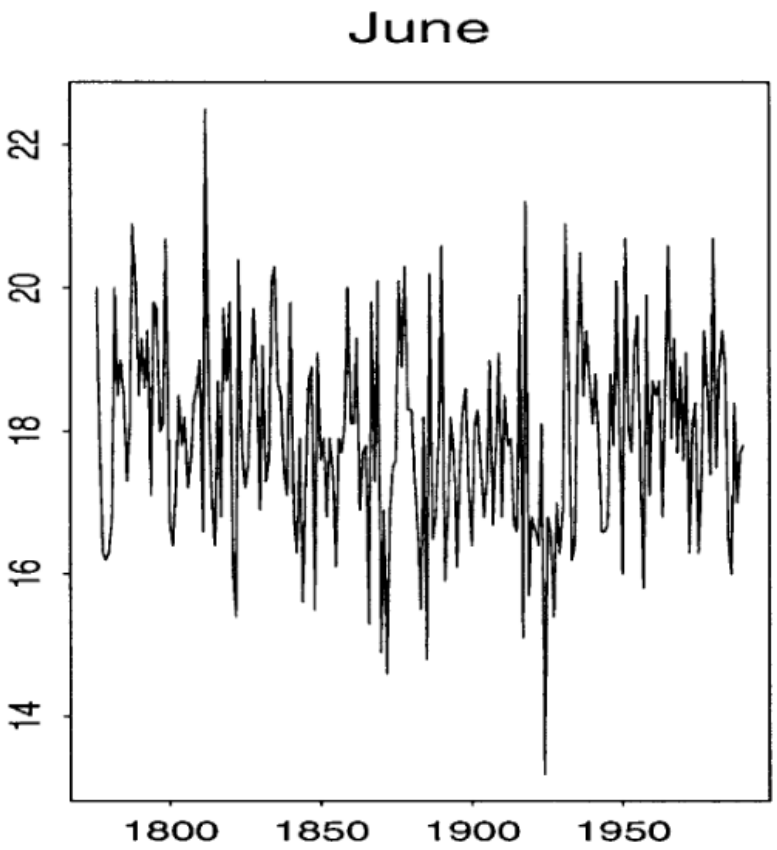


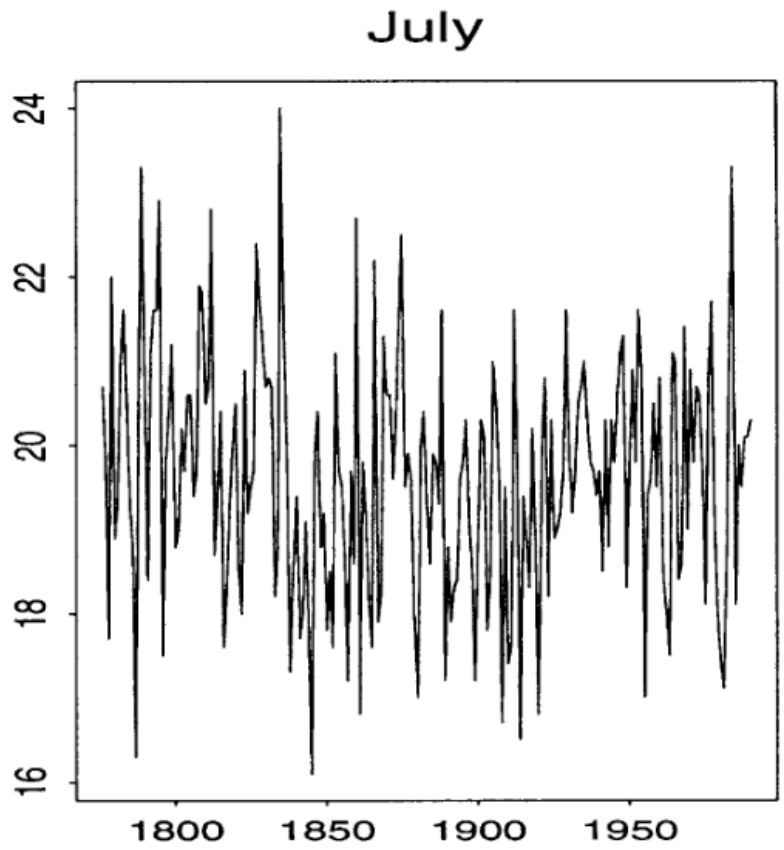
Aplikace na vývoj teploty v Praze (1775 – 1989)

Year of the first change-point detection for the Prague temperature data

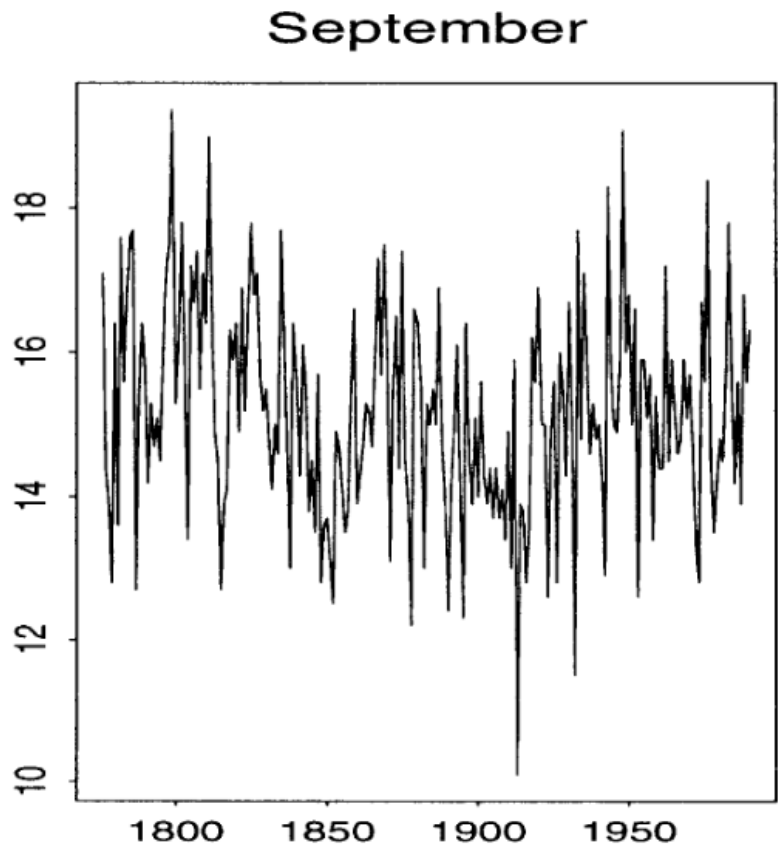
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$D_{0.00}$	NC	1846	NC	NC	1859	1929	1910	1867	1861	NC	1861	NC
$D_{0.25}$	NC	1842	NC	NC	1856	1927	1910	1859	1855	NC	1859	NC
$D_{0.45}$	NC	1830	NC	NC	1856	NC	1917	1865	1855	NC	1859	NC
$\tilde{D}_a$	NC	NC	NC	NC	1862	1928	1914	1884	1895	NC	1877	NC

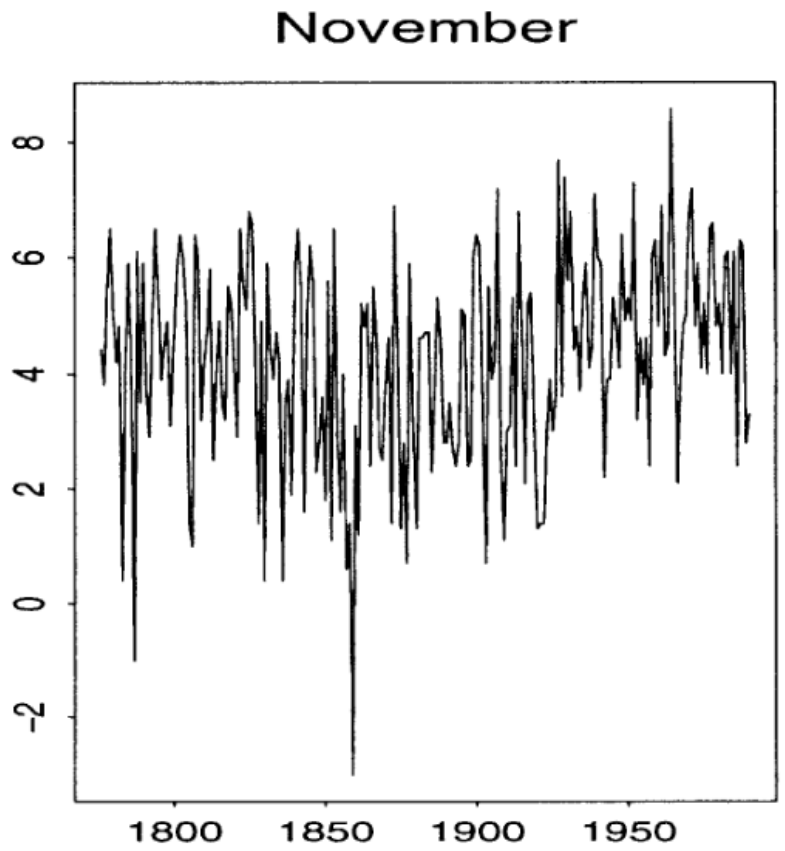
Aplikace na vývoj teploty v Praze (1775 – 1989)





Aplikace na vývoj teploty v Praze (1775 – 1989)





Zdroje

- Horváth, L., Hušková, M., Kokoszka, P., Steinebach, J. (2004). Monitoring changes in linear models. Journal of statistical Planning and Inference, 126(1), 225-251.
- Chu, C.-S.J., Stinchcombe, M., White, H., 1996. Monitoring structural change. Econometrica 64, 1045–1065.