Časové řady II

Ondřej Vencálek
Univerzita Palackého v Olomouci
ondrej.vencalek@upol.cz

seminář pro VŠB-TUO
2015-03-20, Ostrava

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Nové kreativní týmy v prioritách vědeckého bádání

CZ.1.07/2.3.00/30.0055

Tento projekt je spolufinancován z ESF a státního rozpočtu ČR.

Co nás čeká

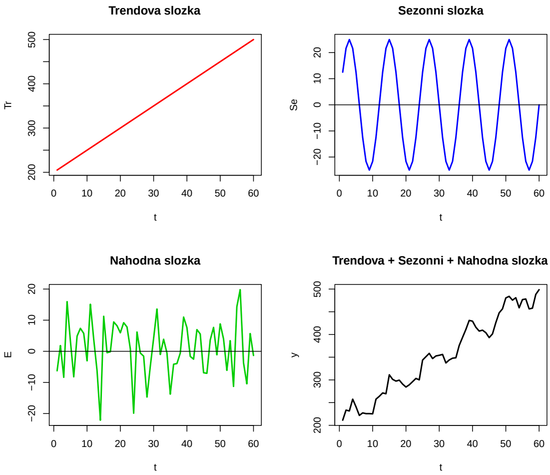
V první části **modelování sezónnosti**

- model konstantní sezónnosti a model proporcionální sezónnosti
- Holt-Wintersova metoda
- model skrytých period

Ve druhé části **úvod do Box-Jenkinsonovy metodologie**

- stacionarita, autokorelační a parciální autokorelační funkce
- procesy MA, AR, ARMA
- procesy ARIMA, SARIMA

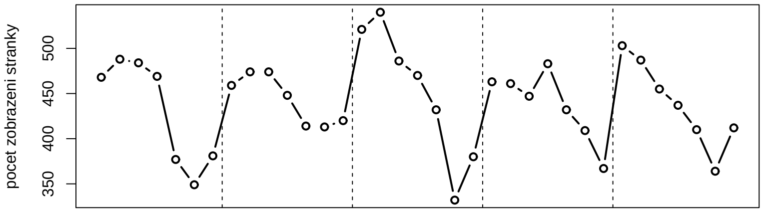
Připomenutí: rozklad č.ř. na jednotlivé složky



Aditivní dekompozice časové řady:

$$y(t) = Tr(t) + Se(t) + E(t)$$

Týdenní periodicita počtu zobrazení jisté stránky wikipedie



Model konstantní sezónnosti – indexy

Indexování času: (i, j), kde

- i je číslo sezóny
- j je číslo pozorování v rámci i-té sezóny

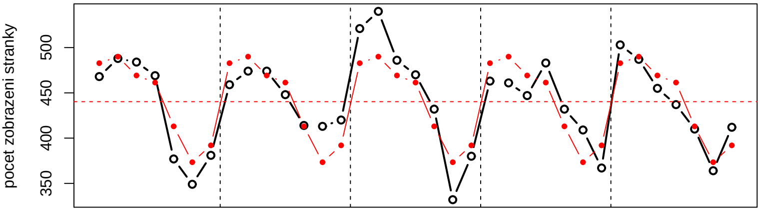
Např. při týdenní periodicitě je pozorování s číslem 11 číslováno dvojicí (2,4), neboť $11 = 1 * 7 + 4$ (jeden celý týden uběhl, nyní je 4. den 2. týdne.

Model konstantní sezónnosti

$$
\begin{aligned}
y_{(i,j)} &= Tr_{(i,j)} + Sz_{(i,j)} + E_{(i,j)} \\
Sz_{(i,j)} &= \beta_j
\end{aligned}
$$

tj. efekt sezónnosti je stejný (konstantní) ve všech sezónách.
Trend může být konstantní, lineární, kvadratický, ...

Model konstantní sezónnosti (s konstantním trendem)



Model:

$y_{i,j} = \alpha + \beta_j + E_{(i,j)}$

Identifikovatelnost

Představme si délku periody 4.
Následující situace jsou „nerozlišitelné“:

- $\alpha = 10$
  $\beta_1 = 1, \beta_2 = 3, \beta_3 = -1, \beta_4 = -1$
- $\alpha = 10,5$
  $\beta_1 = 0,5, \beta_2 = 2,5, \beta_3 = -1,5, \beta_4 = -1,5$
- $\alpha = 11$
  $\beta_1 = 0, \beta_2 = 2, \beta_3 = -2, \beta_4 = -2$
- ...

Musíme si „vybrat“.
Běžně se používá jedna z podmínek identifikovatelnosti:

- $\beta_1 = 0$ (default v R-ku)
- $\sum \beta_j = 0$

Model konstantní sezónnosti (s konstantním trendem) – odhady parametrů

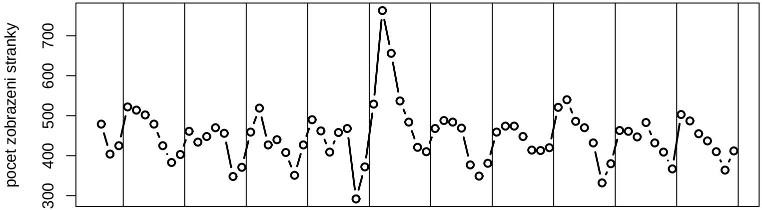
- MNČ
- R-ko: `lm(navstevnost~as.factor(denvtydnu))`

parametr	odhad	p-value
α	482,8	$< 0,001$
β_2	7,2	$0,662$
β_3	-13,6	$0,411$
β_4	-21,4	$0,200$
β_5	-69,8	$< 0,001$
β_6	-109.4	$< 0,001$
β_7	-90.8	$< 0,001$

Ne vždy je situace tak jednoduchá

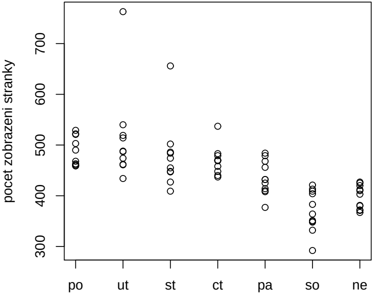
Tatáž časová řada v delším horizontu (od začátku roku 2015 do současnosti). Návštěvnost stránky

http://en.wikipedia.org/wiki/Milan_Kundera



Odlehlá pozorování

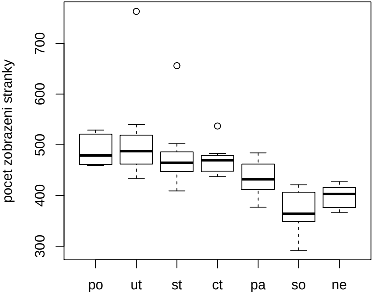
Co je cílem analýzy? Pokud je to odhad efektu dne (v běžné situaci), pak nám outliery komplikují situaci, protože vychylují odhad pro některé dny.



Nějaká vysvětlení outlierů?

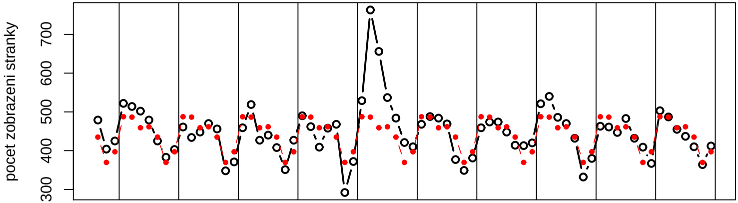
Odlehlá pozorování

Co je cílem analýzy? Pokud je to odhad efektu dne (v běžné situaci), pak nám outliery komplikují situaci, protože vychylují odhad pro některé dny.



Nějaká vysvětlení outlierů?

Výsledné proložení



... bylo získáno odhadem efektů jednotlivých dnů na datech, z nichž byly vyloučeny tři odlehlé hodnoty (tři po sobě jdoucí dny).

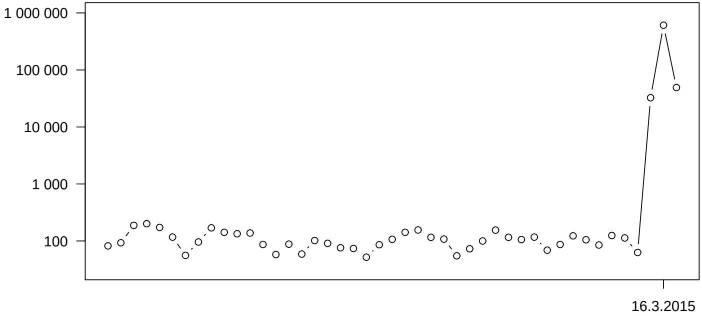
Anna Atkins

Anna Atkins (16. března 1799 – 9. června 1871)
anglická botanička a fotografka; je často považována za první osobu, která vydala knihu ilustrovanou fotografickými obrázky (1843: *Photographs of British Algae: Cyanotype Impressions*)

`http://cs.wikipedia.org/wiki/Anna_Atkinsová`
`http://en.wikipedia.org/wiki/Anna_Atkins`

`https://www.google.com/doodles`

Efekt doodlu



Model proporcionální sezónnosti

$$\begin{aligned} y_{(i,j)} &= Tr_{(i,j)} + Sz_{(i,j)} + E_{(i,j)} \\ Sz_{(i,j)} &= c_j \cdot Tr_{(i,j)} \\ y_{(i,j)} &= (1 + c_j) Tr_{(i,j)} + E_{(i,j)} \end{aligned}$$

tj. efekt sezónnosti je multiplikativní tj. jeho velikost je úměrná velikosti trendu.
Trend může být konstantní, lineární, kvadratický, ...

$$\sum_j c_j = 0$$

Holt-Winters (HW)

- adaptivní přístup k trednové složce
- ukážeme HW pro lokálně lineární trend
- přidána složka sezónní
- multiplikativní a aditivní verze HW
- podstata HW: „dvojí optika“

Holt-Winters

- $y_t = Tr_t + Sz_t + E_t$ (aditivní HW)
  $y_t = Tr_t \cdot Sz_t \cdot E_t$ (multiplikativní HW)
- $Tr_t = \beta_0 + \beta_1 t$

Holt-Winters: rekurentní vztahy pro odhady Tr_t a Sz_t

Situace:

- Pro všechny časy minulé, tj. $\tau = 1, \ldots, t-1$ máme $\widehat{Tr}(\tau)$ tedy $\widehat{\beta}_0(\tau)$ a $\widehat{\beta}_1(\tau)$, a $\widehat{Sz}(\tau)$.
- Chceme odhad jednotlivých složek (Tr_t, Sz_t, β_1) pro současnost, tj. pro čas t, v němž máme nové pozorování y_t .

Dvojí optika: trend

- $y_t \approx Tr_t + Sz_t$
  $Tr_t \approx y_t - Sz_t$
  (pohled využívající novou informaci y_t)
- $Tr_t = \beta_0 + \beta_1 t = \beta_0 + \beta_1(t-1) + \beta_1$
  $Tr_t = Tr_{t-1} + b_1$
  (pohled využívající starých informací)

Využití:

$$\widehat{Tr}(t) = \alpha \left[y_t - \widehat{Sz}(t-L) \right] + (1-\alpha) \left[\widehat{Tr}(t-1) + \widehat{\beta}_1(t-1) \right]$$

Dvojí optika: směrnice trendu

- $Tr_t = \beta_0 + \beta_1 t = \beta_0 + \beta_1(t-1) + \beta_1 = Tr_{t-1} + \beta_1$
  $\beta_1 = Tr_t - Tr_{t-1}$
  (pohled využívající novou informaci)
- trend je lokálně lineární, tedy
  (β_1 v čase t) $\approx$ (β_1 v čase $t-1$)
  (pohled využívající starých informací)

Využití:

$$\widehat{\beta_1}(t) = \xi \left[\widehat{Tr}(t) - \widehat{Tr}(t-1)\right] + (1-\xi)\widehat{\beta_1}(t-1)$$

Dvojí optika: sezónní složka

- $y_t \approx Tr_t + Sz_t$
  $Sz_t \approx y_t - Tr_t$
  (pohled využívající novou informaci y_t)
- při délce sezóny L
  $Sz_t \approx Sz_{t-L}$
  (pohled využívající starých informací)

Využití:

$$\widehat{Sz}(t) = \gamma \left[y_t - \widehat{Tr}(t) \right] + (1-\gamma)\widehat{Sz}(t-L)$$

Váhy α, ξ, γ

- hodnoty v intervalu (0,1)
- jakou váhu přikládáme novým informacím a jakou starým
- minulé odhady v sobě nesou informace z delší minulosti, váhy tedy určují, jak rychle „zapomínáme minulost“
- můžeme nastavit tak, abychm minimalizovali chybu predikcí (viz minule)

Počáteční odhady

Pořebujeme

- $\widehat{Tr}_0(0) = \widehat{\beta}_0(0)$,
- $\widehat{\beta}_1(0)$
- $\widehat{Sz}(1-L), \widehat{Sz}(2-L), \ldots, \widehat{Sz}(L-L)$

Jak si je opatřit?

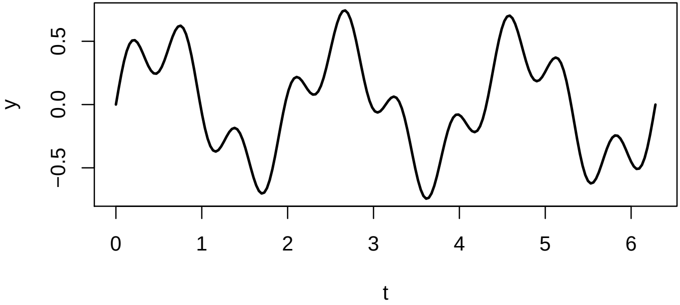
Třeba odhadem parametrů modelu konstantní sezónnosti s lineárním trendem na počátečním úseku č.ř.

Predikce pomocí HW

$$\widehat{y}_{t+\tau}(t) = \widehat{Tr}(t) + \widehat{\beta}_1(t) \cdot \tau + \widehat{Sz}(t + \tau - L)$$

Pokud $\tau > L$, použijeme místo $\widehat{Sz}(t + \tau - L)$ nejaktuálnější odhad sezónní složky odpovídající sezónnosti v čase $t + \tau$, např. $\widehat{Sz}(t + \tau - 2L)$, $\widehat{Sz}(t + \tau - 3L)$, ...

Model skrytých period



$$y_t = 0,5 \cdot \sin(3t) + 0,25 \cdot \sin(10t)$$

Terminologie

$$f(t) = A \sin\left(\frac{2\pi}{T} t + \varphi\right)$$

- A ... amplituda
- T ... perioda
- $\frac{1}{T}$ ... frekvence
- $\frac{2\pi}{T}$ ... úhlová frekvence
- φ ... fázový posun
- t ... čas

Báze prostoru funkcí s periodou T

$$\left\{\sin\left(j\frac{2\pi}{T}t\right), \cos\left(j\frac{2\pi}{T}t\right)\right\}_{j\in\mathbb{N}_0}$$

Model skrytých period

$$y_t = a_0 + \sum_{j=1}^{H} a_j \sin(w_j t) + \sum_{j=1}^{H} b_j \cos(w_j t)$$

- $w_j = \frac{2\pi}{n} j$
- $H = (n-1)/2$ pro n liché, $H = n/2$ pro n sudé
- n je základní perioda, $n/2, n/3, \ldots, 2$ vedlejší periody
- $a_0, a_1, \ldots, a_H,\ b_1, \ldots, b_H$ neznámé parametry

Odhad parametrů

MNČ

$$
\begin{aligned}
\widehat{a}_0 &= \frac{1}{n}\sum_{t=1}^{n} y_t = \bar{y} \\
\widehat{a}_k &= \frac{2}{n}\sum_{t=1}^{n} y_t \sin(w_k t) \quad k = 1, \ldots, H \\
\widehat{b}_k &= \frac{2}{n}\sum_{t=1}^{n} y_t \cos(w_k t) \quad k = 1, \ldots, H
\end{aligned}
$$

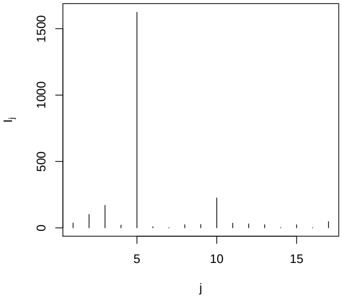
Periodogram

$$\sum_{t=1}^{n}(\hat{y}_t - \bar{\hat{y}}_t)^2 = \frac{1}{2}\sum_{j=1}^{H}(\hat{a}_j^2 + \hat{b}_j^2)$$

Označme

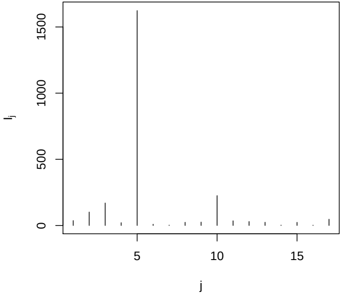
$$I_j = (\hat{a}_j^2 + \hat{b}_j^2), \quad j = 1, \dots, H$$

Čím větší I_j, tím více j-tá perioda přispívá k celkové variabilitě
Periodogram pro data o návštěvnosti stránky M.Kundery na wikipedii:



Periodogram – příklad

Periodogram pro data o návštěvnosti stránky M.Kundery na wikipedii:



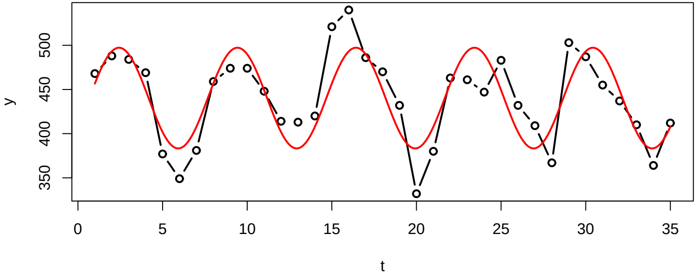
$n = 35$ (denní data za 5 týdnů)

$$w_5 = \frac{2\pi}{35} 5 = \frac{2\pi}{7},$$

tedy délka nejvýznamnější periody je 7 dní čili týden

Příklad – model s jedinou významnou frekvencí

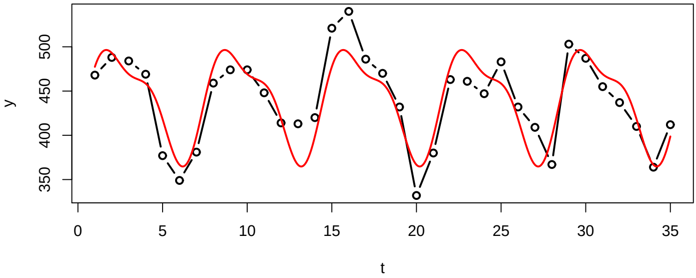
$$y_t = 440,3 + 46,9 \sin\left(\frac{2\pi}{7} t\right) - 32,4 \cos\left(\frac{2\pi}{7} t\right)$$



Příklad – další významné frekcence

- $k = 10 \Rightarrow T = 3,5 = 7/2$: přidáním této frekvence není porušena týdenní periodocita
- $k = 3 \Rightarrow T = 11,67$: perioda, která poruší týdenní periodicitu – úrovně různých týdnů budou různé; dlouhodobější kolísání

Příklad – model frekvencí 1/7 a 2/7



Příklad – model frekvencí 1/7 a 2/7 a 3/35