

Analýza časových řad

Základní informativní přehled

Dekompozice ČŘ

Box-Jenkinsovy modely (AR, MA, ARMA)

Spektrální analýza

Časová řada (time serie)- definice

- **Časová řada (ČŘ)** je chronologicky uspořádaná posloupnost pozorování určitého statistického ukazatele.
- Jedná se o spojitý znak měřený v pravidelných (nepravidelných) časových intervalech
- Diskrétní body pozorování $t=1, \dots, n$:

$$y_i, i = 1, \dots, n$$

$$y_1, y_2, y_3, \dots, y_n$$

Příklady časových řad

- Vývoj ekonomických ukazatelů (burza, pasažéři přepravení leteckou společností)
 - Průtok řek (Nil)
 - Aktivita slunce – sluneční skvrny (perioda 11 let)
 - Tělesná teplota organismu
-

Úlohy řešené v souvislosti s ČŘ

- **Porozumění mechanismu**, na základě kterého jsou generovány sledované údaje. (př. rozpoznání cyklického chování, určení trendu)
 - Znalost modelu umožňuje **predikování** budoucího vývoje.
 - *Poznámka:* budeme zkoumat statistickou časovou řadu (její chování je zatíženo nejistotou). Nebudeme zkoumat deterministické časové řady (chování lze striktně popsat matematickými vzorci).
-

Specifické problémy časových řad

- Problémy s **volbou časových bodů** pozorování
(atmosférický tlak)
 - Problémy s **kalendářem**
př. nestejný počet dní v měsíci
 - **Nesrovnatelnost** jednotlivých měření
chybí měření na některých subjektech
změna způsobu měření – hladina řeky NIL
 - Nedostatečná **délka** časové řady
příliš krátká – nedoporučují se některé metody
příliš dlouhá – mění se charakteristiky řady
-

Složky časových řad

- **Trend** (odráží dlouhodobý vývoj) - Tr_t
 - **Sezónní složka** - seasonal component (periodická složka se známou periodou, př. rok, den) - Sz_t
 - **Cyklická složka** (periodická složka s neznámou periodou) - C_t
 - **Residuální složka** (zbytková, náhodná, iregulární), E_t
na rozdíl od regrese nejsou chybové složky mezi sebou nezávislé – *autokorelace!* – všechny vlivy, které na časovou řadu působí a které nedokážeme systematicky podchytit a popsat
-

Náhodná složka: platí –li předpoklady

1. $E(\varepsilon_i) = 0$ pro každé $i=1,2,\dots,n$
Střední hodnota náhodné složky je nulová. Tato podmínka znamená, že náhodná složka nepůsobí systematickým způsobem na hodnoty časové řady Y_t .
2. $E(\varepsilon_i^2) = \sigma^2$ pro každé $i=1,2,\dots,n$
Rozptyl náhodné složky je konstantní. Tato podmínka vyjadřuje, že variabilita náhodné složky nezávisí na hodnotách systematických složek a je rovna neznámé kladné konstantě.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ pro každé $i \neq j=1,2,\dots,n$
Kovariance náhodné složky je nulová. Hodnoty náhodné složky jsou tedy nekorelované.
4. ε_i mají normální rozdělení pro každé $i=1,2,\dots,n$.

1+2+3 bílý šum

1+2+3+4 normální bílý šum

Dekompozice časových řad

- Aditivní dekompozice

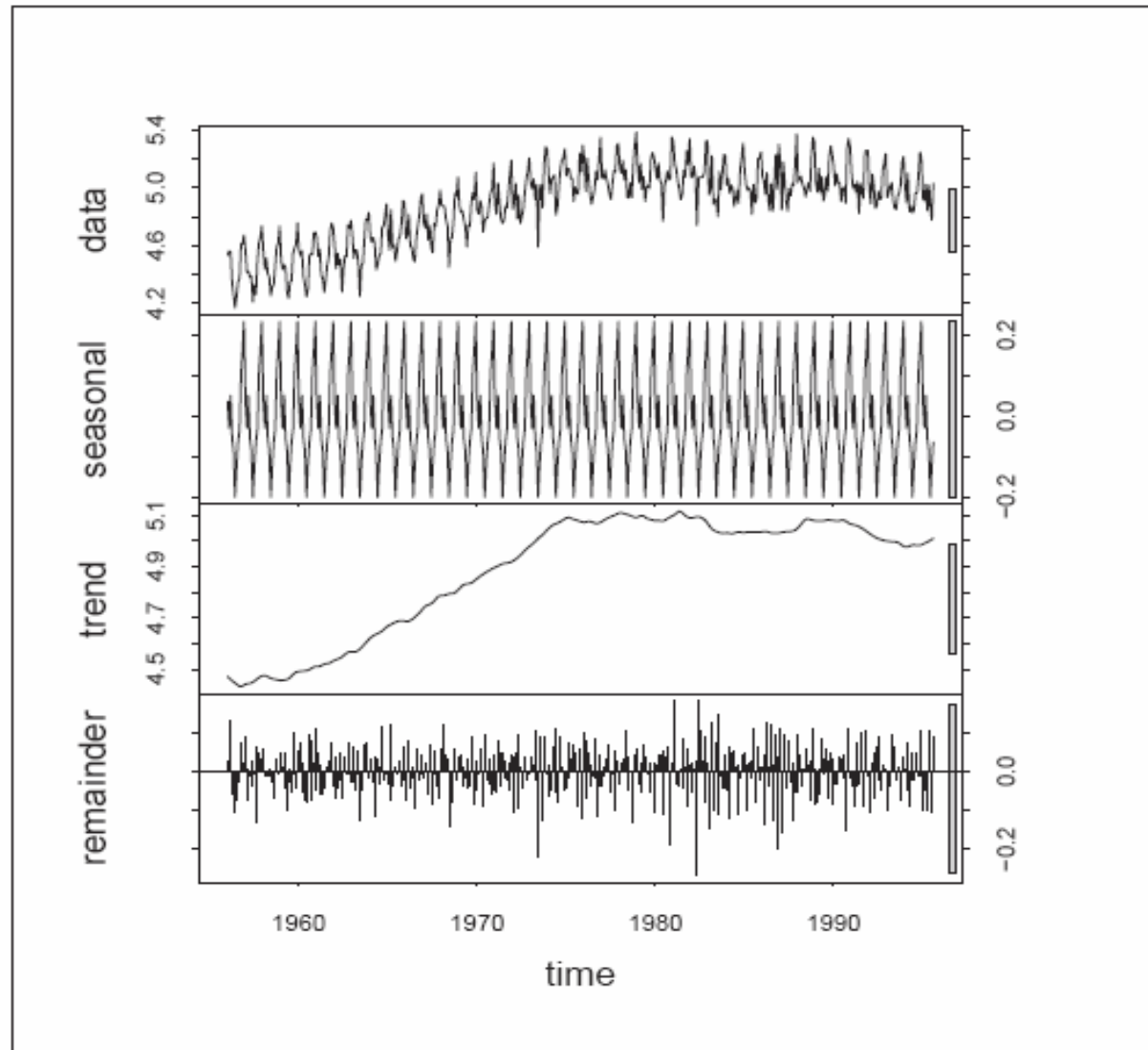
$$y_t = Tr_t + C_t + Sz_t + E_t$$

- Multiplikativní dekompozice

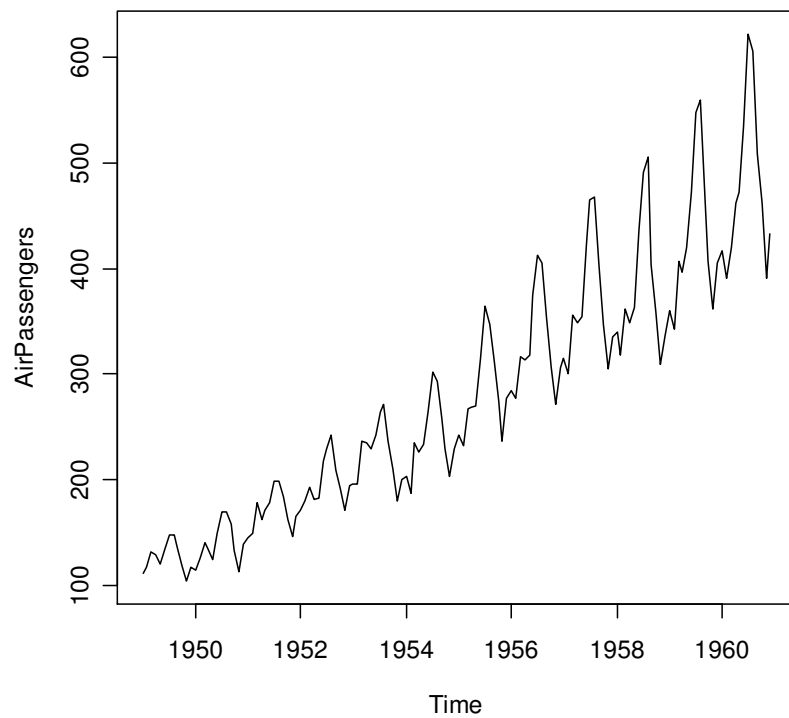
$$y_t = Tr_t \cdot C_t \cdot Sz_t \cdot E_t$$

Aditivní dekompozice

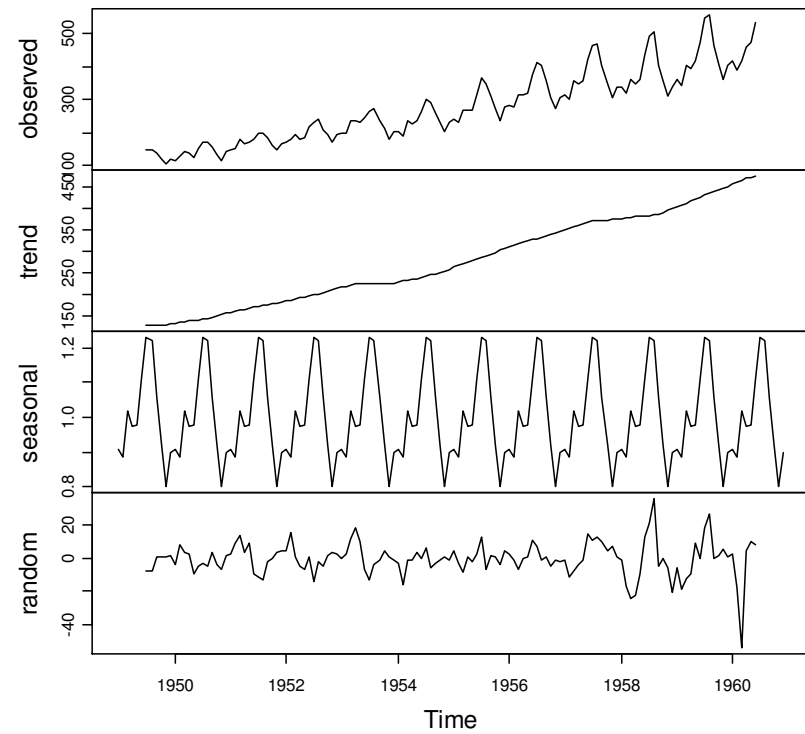
- Měsíční produkce piva – Austrálie leden 1956 - srpen 1995
- Trend – Tt lokální vyhlazení – loess regression
- sezónní složka St (a residua et) vypočteny z rozdílu $Xt - Tt$.



Multiplikativní dekompozice



Decomposition of multiplicative time series



Metody modelování trendu

- Adaptivní postupy modelování trendu
 - klouzavých průměrů
 - exponenciálního vyrovnávání
- Modelování pomocí matematických křivek

Periodická složka a autokorelace

- Lze prokazovat a hodnotit po odstranění trendu a sezónního kolísání – „očištění“ ČŘ *od některých složek, odfiltrování*
 - Speciální postupy – nelze mechanicky využít lineární regresi či lineární vyhlazování
-

Modelování pomocí matematických křivek

- *Tvar časové řady: $Y_t = T_t + \varepsilon_t$*
 - *Odhad parametrů trendové funkce – metoda nejmenších čtverců*
 - *Regresní model, kde v roli vysvětlující proměnné vystupuje čas*
-

Modelování pomocí matematických křivek

- *konstantní trend (někdy se taková řada označuje jako řada bez trendu)*

$$T_t = \beta_0 \quad \text{pro } t=1,2,\dots,n$$

- *lineární trend*

$$T_t = \beta_0 + \beta_1 t \quad \text{pro } t=1,2,\dots,n$$

- *kvadratický trend*

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 \quad \text{pro } t=1,2,\dots,n$$

- *exponenciální trend*

$$T_t = \beta_0 \beta_1^t \quad \text{kde } \beta_1 > 0, \quad \text{pro } t=1,2,\dots,n$$

- *modifikovaný exponenciální trend*

$$T_t = \beta_0 + \beta_1 \beta_2^t \quad \text{kde } \beta_2 > 0, \quad \text{pro } t=1,2,\dots,n$$

- *logistický trend*

$$T_t = \gamma / (1 + \beta_0 \beta_1^t) \quad \text{kde } \beta_2 > 0, \quad \text{pro } t=1,2,\dots,n$$

- *Gompertzova křivka*

$$\ln(T_t) = \gamma + \beta_0 \beta_1^t \quad \text{kde } \beta_2 > 0, \quad \text{pro } t=1,2,\dots,n$$

Metoda klouzavých průměrů

(moving averages, MA)

- **Adaptivní přístup** k trendové složce
- Adaptivní metody jsou schopny pracovat s trendovými složkami, které se mění v čase.
- Pozorování Y_t porovnáváme s váženým průměrem (lineární kombinací) m sousedních pozorování (včetně Y_t). Např.

$$\frac{1}{35}(-3Y_{t-2} + 12Y_{t-1} + 17Y_t + 12Y_{t+1} - 3Y_{t+2})$$

- Snaha vyhladit nahodilé výchyly a zachovat „průměrný“ vývoj
 - **Lokální vyrovnávání trendu, koncepce postupného trendu**
 - Vhodné zejména k interpolaci a nalezení dosud přehlížených systematických vlivů - **Vyrovnané hodnoty mají funkci trendu**
 - Jednoduché klouzavé průměry – běžné aritmetické průměry.
-

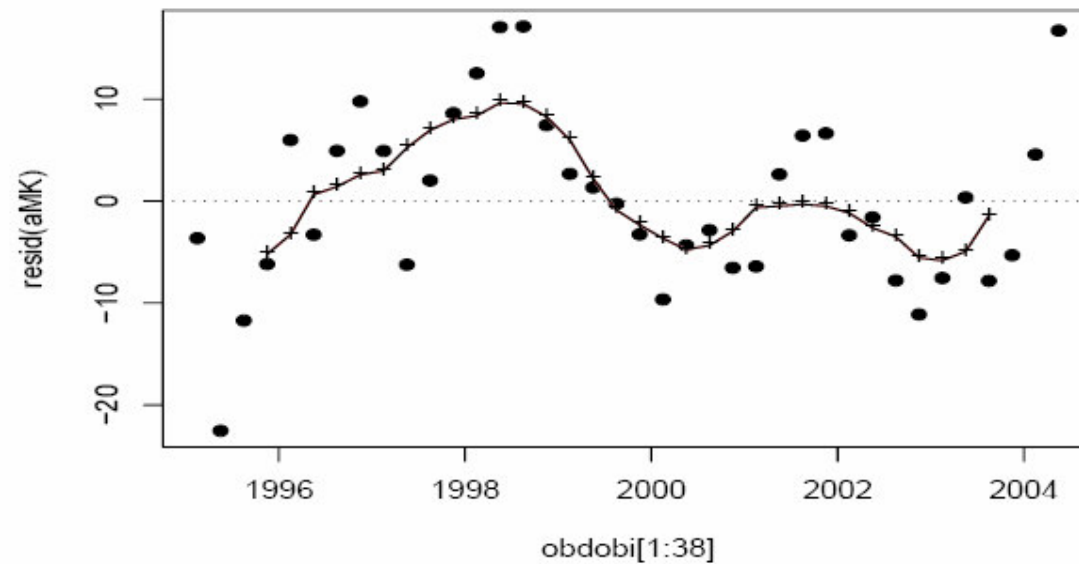
Metoda klouzavých průměrů

(moving averages, MA)

- Určení vah
 1. Koeficienty(váhy) pevně dané a konstantní pro celou čas.řadu nebo odhadnuty viz.bod 2.
 2. MA jsou ekvivalentní vyrovnávání krátkých úseků řady určitými matematickými křivkami -> odhad parametrů.
 - Vlastnosti:
 1. Součet vah klouzavých průměrů je roven 1.
 2. Váhy jsou symetrické kolem střední hodnoty
 - Problémy:
 1. volba řádu matematické křivky
 2. volba délky posloupnosti, m – (čím větší délka, tím větší vyrovnání)
-

Zvára: Statistika

příklad: klouzavé průměry reziduí ($m=7$)



Exponenciální vyrovňávání (exponential smoothing)

- **Adaptivní přístup** k trendové složce
 - Výchozí idea: v posledním pozorování se projevuje vliv všech dostupných předchozích
 - Tento vliv je postupně utlumován: tj. největší vliv má bezprostředně předcházející pozorování, nejmenší vliv má pozorování v čase nejvzdálenější. (tj. váhy nejsou symetrické)
 - Použitelné k předpovědi
 - Model exponenciálního vyrovňávání – de facto speciální případ Box-Jenkinsových modelů.
-

Exponenciální vyrovnávání

- Modifikace metody nejmenších čtverců
- Váhy jednotlivých čtverců se směrem do minulosti exponenciálně zmenšují.
- Symbolem se stříškou označujeme vyrovnanou hodnotu v čase t . Vzhledem k těmto hodnotám je minimalizován výraz

$$(y_t - \hat{y}_t)^2 + (y_{t-1} - \hat{y}_{t-1})^2 \alpha + (y_{t-2} - \hat{y}_{t-2})^2 \alpha^2 + \dots$$

Kde $0 < \alpha < 1$ je vyrovnávací hodnota.

- Jednoduché (konstantní trend), dvojité (lineární trend) a trojitě (kvadratický trend) exponenciální vyrovnávání.
-

Jednoduché exponenciální vyrovňování

- Výraz pro výpočet exponenciálně vyrovnaných hodnot se častěji převádí do následujícího (rekurentního) tvaru:

- **Praktické využití:**

- Nejprve položíme: $\hat{y}_1 = y_1$

- A dále: $\hat{y}_2 = ay_2 + (1-a)\hat{y}_1$

$$\hat{y}_3 = ay_3 + (1-a)\hat{y}_2$$

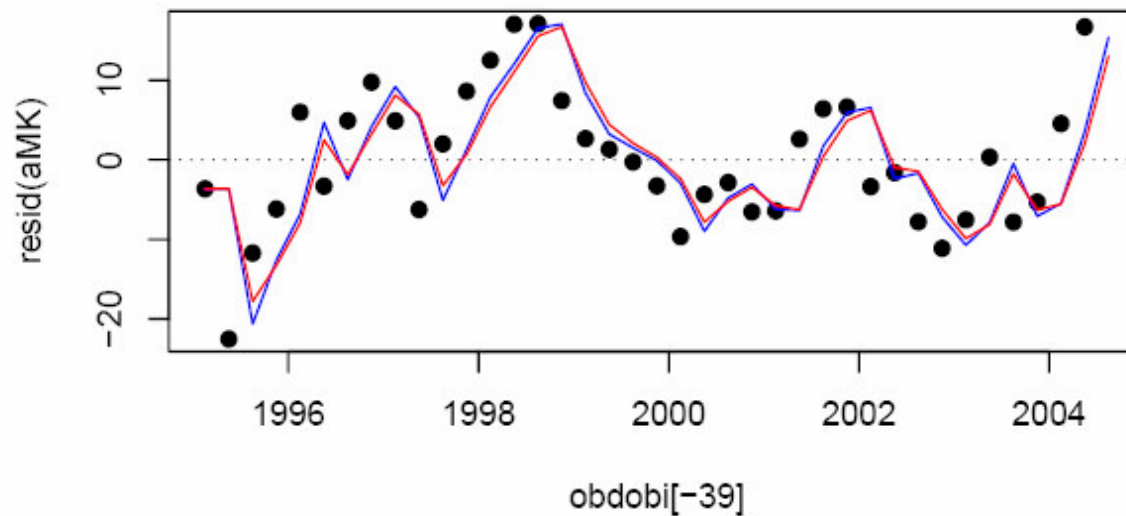
- Obecně: $\hat{y}_t = ay_t + (1-a)\hat{y}_{t-1}$

- **Interpretace:**

Exponenciálně shlazená hodnota v čase t je rovna váženému součtu hodnoty řady v tomto čase t (s vahou a .) a předchozí shlazené hodnoty v čase $t-1$ (s vahou $a-1$).

Zvára: Statistika

exponenciální vyrovnávání reziduí (704,0 pro $w=0,1$, 701,6 pro $w=0,25$)



Box-Jenkinsovy modely

- **Dekompoziční metody** – práce se systematickými složkami ČŘ (trend, sezónní a cyklická složka), jednotlivá pozorování se berou jako nekorelovaná.
 - **Box-Jenkinsovy modely** – základní prvek konstrukce modelu ČŘ je residuální složka, která je tvořena korelovanými (závislými) náhodnými veličinami. Zaměřuje se na vyšetření závislosti – korelační analýza.
-

Typy modelů

- Modely pro residuální, náhodnou složku
 1. Model klouzavých součtů (*Moving average*), MA
Nezaměňovat s metodou klouzavých průměrů
 2. Autoregresní modely, AR
 3. Smíšené modely, ARMA
 - Modely pro ČŘ se zjevným trendem nebo sezónní složkou
 1. Integrované modely, ARIMA
-

Výhody a nevýhody

- Tvorba modelu přímo z dat - nelze vytvořit model „hypoteticky“, na základě nějaké teorie
 - Výhody
 1. Flexibilita, rychlá adaptace na změny v charakteru ČŘ
 2. Dobré praktické výsledky
 - Nevýhody
 1. Požadovaná délka ČŘ je alespoň 50 pozorování.
 2. Neexistuje jednoduchá interpretace. Řada je modelována pomocí náhodných šoků. 😊
-

Stacionarita časové řady

- **Stacionarita** znamená, že chování časové řady je stochasticky ustálené.
- **Striktní stacionarita** – pravděpodobnostní chování je invariantní vůči posunům čase. Pro libovolné h je rozdělení vektoru $(y_{t_1}, \dots, y_{t_k})$ stejné jako rozdělení vektoru $(y_{t_1+h}, \dots, y_{t_k+h})$
- **Slabá stacionarita** – konstantní střední hodnota, konstantní rozptyl, kovariance invariantní vůči posunu v čase, tj. pro libovolné h platí

$$\text{cov}(y_t, y_s) = \text{cov}(y_{t+h}, y_{s+h})$$

Závislost mezi dvěma libovolnými pozorováními závisí na vzájemném časovém umístění, nikoliv na jejich skutečném časovém umístění.

-
- Pomocí Box-Jenkinsovy metodologie lze modelovat jen slabě stacionární časové řady.
 - Nestacionární ČŘ lze převést na stacionární pomocí transformací (např. diferenciací).
-

Autokovarianční a autokorelační funkce

- Autokorelační koeficient – jedná je o korelační koeficient vypočtený mezi členy časové řady a stejné časové řady posunuté o k .

$$\text{corr}(Y_t, Y_{t+k})$$

- Autokovarianční funkce (ACF) – závislost mezi hodnotami autokorelačního koeficientu a hodnotami posunu k . Hodnoty autokorelační funkce se pohybují v rozmezí od -1 do 1.
 - Korelogram – grafický záznam autokorelačních koeficientů pro jednotlivá k .
 - **K0** hodnota, za kterou bývá autokorelační funkce nulová
-

Metoda klouzavých součtů

- Nejjednodušší model MA(1)

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

- Model MA(q)

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

Kde ε_t je bílý šum a θ_j jsou parametry

K0=q

Autoregresní proces

- Nejjednodušší model AR(1)

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t$$

- Model AR(p)

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \varepsilon_t$$

Kde ε_t je bílý šum a φ_j jsou parametry

ARMA

- Nejjednodušší model ARMA(1,1)

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

- Model ARMA(p,q)

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

Kde ε_t je bílý šum a φ_j, θ_k jsou parametry

Spektrální analýza

(P. Dobrovolný: Statistické metody zpracování dat)

- založena na předpokladu, že časová řada je součtem funkcí \sin a \cos o rozličných amplitudách a frekvencích.
 - Tato koncepce především umožňuje nalézt významná cyklická kolísání.
 - Pracuje s pojmy **spektrum** a **frekvence**.
 - Analýza vyšetřuje spektrum řady (tj. zjišťuje intenzitu zastoupení jednotlivých frekvencí a jejich parametrů v časové řadě).
 - Předmětem analýzy není časová proměnlivost ale změny ve frekvencích.
-

Frekvence, Perioda

- **FREKVENCE (f)** – počet cyklů realizovaných za jednotku času. Např. počet složenek na poště má frekvenci 12 cyklů za rok tj. $f=12$.
 - **PERIODA (T)** – čas potřebný k realizaci jednoho cyklu $T=1/f$, tedy frekvence 12 představuje periodu $T=1/12 = 0,0833$ roku.
 - Jestliže dosavadní metody analýzy lze označit jako metody v **časové doméně** (oboru), periodická a cyklická kolísání lze dobře studovat v tzv. **spektrální doméně**.
-

Princip spektrální analýzy

$$y_t = a_0 + \sum_{k=1}^q [a_k \cos(2\pi f_k t) + b_k \sin(2\pi f_k t)]$$

$$f_k = k / q$$

- Fourierova řada ($q=\infty$)
-

Princip spektrální analýzy

- Parametry vypovídají o tom, **do jaké míry příslušná funkce *sin* či *cos* koreluje s daty v časové řadě.**
 - Hodnota q označuje počet *sin* či *cos* funkcí, které jsou použity pro rozklad řady.
 - Spektrální analýza identifikuje stupeň korelace funkcí *sin* či *cos* s různou frekvencí s pozorovanými hodnotami časové řady.
 - Vysoká hodnota koeficientu *sin* či *cos* značí, že v dané časové řadě je **silně zastoupena periodická složka s odpovídající frekvencí** (periodou).
-

Interpretace výsledků

- Vysoká hodnota určitého koeficientu tedy říká, že v časové řadě je obsažena významná cykličnost s danou frekvencí (či délkou periody).
 - K interpretaci výsledků rozložení časové řady na jednotlivé sin a cos členy jsou vhodné **grafické metody**.
 - Znázorňují hodnoty „**periodogramu**“ či hodnoty „**spektrální hustoty**“ vypočtené pro jednotlivé frekvence (periody).
-

Periodogram

- Funkce sin a cos jsou vzájemně nezávislé (ortogonální) – potom můžeme vypočítat sumu druhých mocnin koeficientů pro každou frekvenci a obdržet tak hodnotu periodogramu

$$P_k = (a_k^2 + b_k^2) * N / 2$$

- P_k – hodnota periodogramu na frekvenci fk
 - N – počet členů časové řady
 - Neparametrická metoda odhadu spektrální hustoty
-

Spectral plot

- Hodnoty periodogramu obsahují mnoho náhodných fluktuací, mnoho vrcholů.
 - Pro analýzu je podstatnější nalézt takové oblasti frekvencí, které obsahují mnoho sousedních frekvencí.
 - Takových, které nejvíce přispívají k cyklickému chování řady - tedy oblasti frekvencí s největšími spektrálními hustotami.
 - Vyhlazením periodogramu, např. pomocí klouzavých průměrů
-

Literatura

- Cipra T.(1986): Analýza časových řad s aplikacemi v ekonomii, SNTL/Alfa
 - Zvára K.: učební text pro předmět Statistika (D360P03Z, D360P03U)
<http://www.karlin.mff.cuni.cz/~zvara>
 - Dobrovolný P.: Statistické metody zpracování dat – Analýza časových řad, zdroj IS
 - **IASTAT - INTERAKTIVNÍ UČEBNICE STATISTIKY**
<http://iastat.vse.cz/>
 - Časové řady v Rku: <http://www.statোক.wiso.uni-goettingen.de/veranstaltungen/zeitreihen/sommer03/>
-