KGG/STG Statistika pro geografy

12. Analýza časových řad - pokračování

Mgr. David Fiedor
11. května 2015

Verifikace modelu

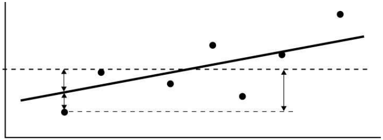
- zhodnocení statistické významnosti odhadnutých parametrů modelu i modelu jako celku
- podstatou metody nejmenších čtverců je fakt, že model vysvětlí vždy pouze část variability pozorovaných dat
- snažíme se zjistit (na dané hladině významnosti), zda model jako celek dává lepší vysvětlení než je možné očekávat jako důsledek náhody
- základním ukazatelem vhodnosti modelu je *koeficient determinace* + *analýza rozptylu*

Analýza rozptylu

A) rozptyl celkový $s_y^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$

B) rozptyl vyrovnaných hodnot (*modelový*) $s_{\hat{y}}^2 = \frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_t)^2$

C) rozptyl reziduální $s_{y-\hat{y}}^2 = \frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_t)^2$



Kritéria pro volbu vhodného modelu trendové funkce

- spočívají v minimalizaci předem zvoleného kritéria (obdobně jako u regresní analýzy)
- kritériem nejčastěji bývá součet čtverců odchylek empirických hodnot y_t od hodnot vyrovnaných $\hat{y}_t$
- konkrétní kritéria:
    - součet čtvercových chyb (SSE)
    - střední chyba odhadu (ME)
    - střední čtvercová chyba odhadu (MSE)

Součet čtvercových chyb

$$SSE = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

- jedná se o formální kritérium (při použití např. polynomu vysokého stupně můžeme získat i nulový reziduální součet čtverců, avšak model bude zcela nepoužitelný)

Střední chyba odhadu

- obvykle v nabídce statistického software
- M.E. (Mean Error)

$$M.E. = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)}{n}$$

Střední čtvercová chyba odhadu

- obvykle v nabídce statistického software
- M.S.E. (Mean Square Error)
- nejpoužívanější kritérium

$$M.S.E. = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}$$

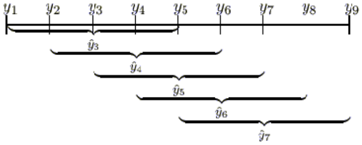
Informativní testy

Pro volbu vhodné trendové křivky:

Trend	**Informativní test**
lineární	První diference (y_{t+1} - y_t) jsou přibližně konstantní
kvadratický	Druhé diference (y_{t+2} - $2y_{t+1}$ + y_t) jsou přibližně konstantní
exponenciální	Podíly sousedních hodnot (y_{t+1}/y_t) resp. První diference logaritmů tvaru ($\log y_{t+1}$ - $\log y_t$) jsou přibližně konstantní
logistický	Křivka prvních diferencí (y_{t+1}-y_t) se podobá křivce normální hustoty, podíly ($1/y_{t+2}$ - $1/y_{t+1})/(1/y_{t+1}$ - $1/y_t$) jsou přibližně konstantní

Mechanické vyrovnávání trendu metodami klouzavých průměrů

- v případě, že trend se mění a nelze ho vyrovnat jednou křivkou
- metoda vhodná pro neperiodické řady



- prosté i vážené
- centrované i necentrované

Klouzavé průměry

Prostý klouzavý průměr

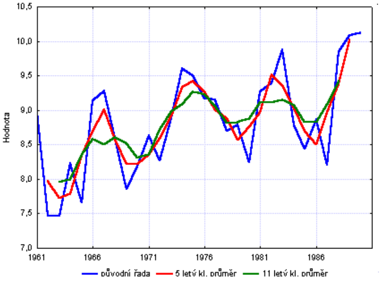
$$\frac{1}{5}(y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})$$

Vážený klouzavý průměr

$$\frac{1}{8}(y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2})$$

Volba řádu klouzavých průměrů

- subjektivní posouzení charakteru dat
- délka klouzavých průměrů by měla odpovídat periodě sezónních či cyklických změn
- obsahuje-li řada sezónní složku, je vhodné volit řád klouzavých průměrů tak, aby zahrnoval celou délku periody sezónní složky



Centrované klouzavé průměry

- především se využívají klouzavé průměry lichého řádu
- u časových řad s klouzavým průměrem sudého řádu využijeme „centrování“ - průměr dvou sousedních klouzavých průměrů lichého řádu

Příklad

U časové řady teplot vzduchu chceme zvolit shlazení klouzavými průměry řádu 12.

Řešení

- nutno vypočítat průměr dvou sousedních klouzavých průměrů liché délky, tedy:

$$\hat{y} = \frac{1}{2}\left(\frac{1}{12}(y_{t-6} + y_{t-5} + \cdots + y_{t+5}) + \frac{1}{12}(\frac{1}{12}(y_{t-5} + y_{t-4} + \cdots + y_{t+6})\right) = \frac{1}{24}(y_{t-6} + 2y_{t-5} + 2y_{t-4} + \cdots + 2y_{t+5} + y_{t+6})$$

Analýza sezónní složky časových řad

1. klasický přístup k sezónní dekompozici
2. úvod do autokorelační analýzy

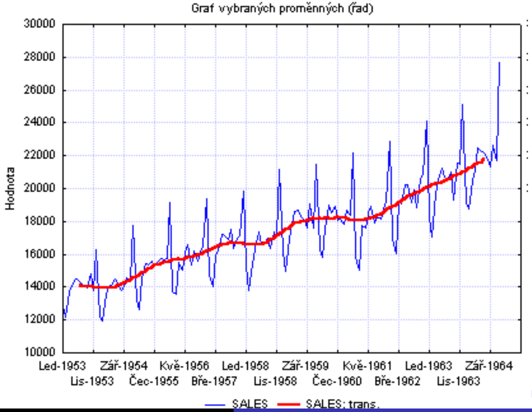
- sezónní složka je typická pro časové řady, jejichž interval pozorování je kratší než jeden rok (týden, měsíc, roční období)
- jak v časových řadách ekonomických (tržby, produkce,. . .), tak např. meteorologických (teplota vzduchu,. . .)

Obecný model řady při sezónním „očišťování“

- trendovou a cyklickou složku považujeme za jeden celek, tedy:
$$Y_t = TC_t + S_t + \epsilon_t,$$
kde Y_t je pozorovaná hodnota časové řady v čase t

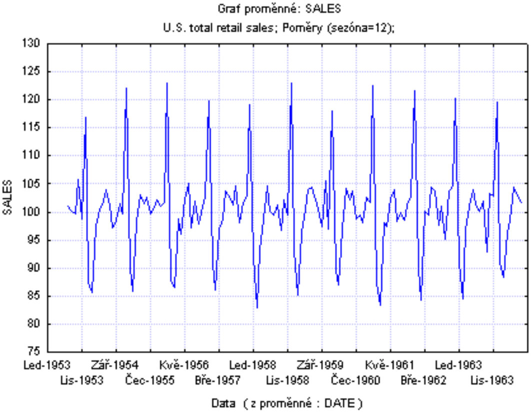
Jednotlivé kroky analýzy sezónní složky

1) Z originální řady obsahující sezónní složku je vypočtena řada klouzavých průměrů s řádem klouzavých průměrů rovným délce sezónní složky:



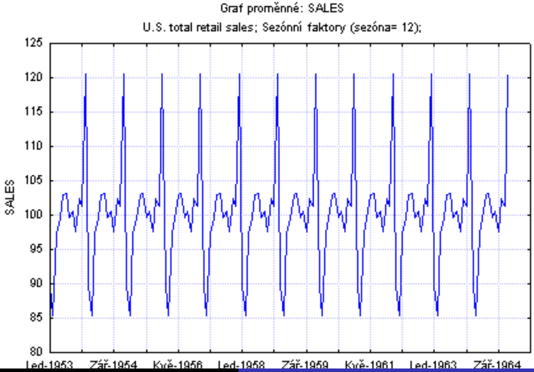
Jednotlivé kroky analýzy sezónní složky

2) Vytvoření nové řady odečtením řady shlazené od řady původní:



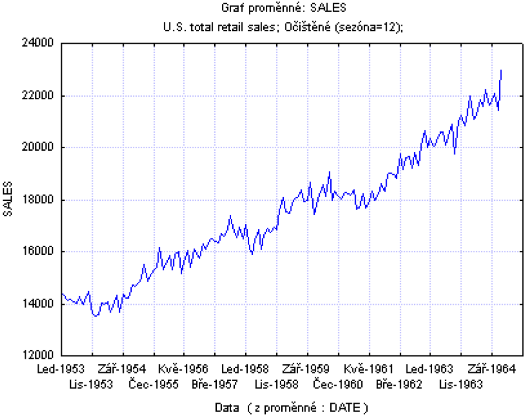
Jednotlivé kroky analýzy sezónní složky

3) Sezónní komponenty jsou vypočteny jako průměr pro každý člen v rámci sezóny (výsledné hodnoty představují průměrnou sezónní složku v časové řadě):



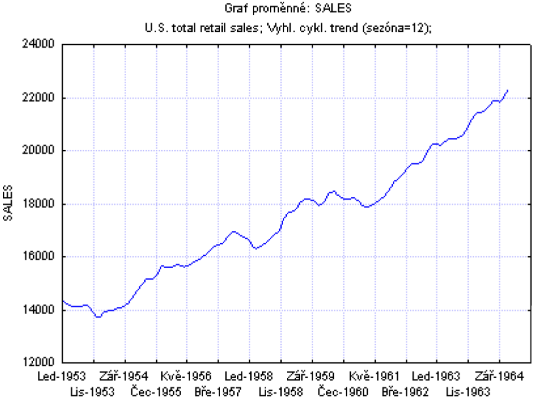
Jednotlivé kroky analýzy sezónní složky

4) „Sezónně očištěná řada“ se vyjádří jako rozdíl řady originální a sezónní komponenty:



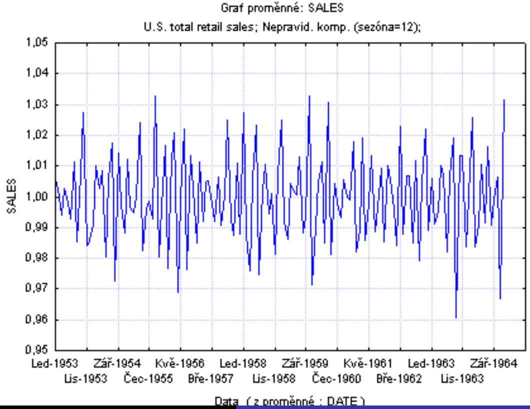
Jednotlivé kroky analýzy sezónní složky

5) Složka TC_t se většinou aproximuje řadou shlazenou váženým klouzavým průměrem řádu 5 se symetrickými vahami (1, 2, 3, 2, 1):



Jednotlivé kroky analýzy sezónní složky

6) Obdobně lze také izolovat náhodnou složku jako rozdíl řady sezónně očištěné a řady se zvýrazněnou složkou TC_t:



Autokorelace časových řad

- metoda, kterou lze zkoumat vzájemné vztahy mezi hodnotami jedné časové řady
- může se využít jako metoda k definování sezónní a cyklické složky časových řad
- základem je výpočet autokorelačního koeficientu, resp. autokorelační funkce

Autokorelační koeficient r_k

- relativní míra proměnlivosti členů časové řady posunutých o určitou hodnotu k
- definuje vztah mezi členy časové řady y_t a y_{t+k}
- posun k se z angličtiny označuje jako *lag*
- jedná se o korelační koeficient vypočtený mezi jednotlivými členy časové řady, mezi kterými je $k-1$ jiných pozorování a označujeme jej jako autokorelační koeficient k-tého řádu

Zavedené pojmy

- rozptyl (variance) - míra variability statistického znaku x
- kovariance - absolutní míra vzájemné variability dvou statistických znaků $x; y$
- korelace - relativní míra vzájemné variability dvou statistických znaků $x; y$

Základní vztahy

- autokovariance - absolutní míra proměnlivosti členů časové řady y posunutých o určitou hodnotu k

$$c_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{n-k-1}$$

- autokorelace - relativní míra proměnlivosti členů časové řady y posunutých o určitou hodnotu k

$$r_y(k) = \frac{c_k}{c_0} = \frac{c_k}{s_y^2}$$

Autokorelační funkce

- závislost mezi hodnotami autokorelačního koeficientu a hodnotami posunu k
- vyjadřuje se formou grafu - korelogramu (na ose x jsou hodnoty $\text{lag}(k)$, na ose y hodnoty autokorelačního koeficientu)
- vhodný nástroj k posouzení, zda časová řada obsahuje cyklickou či periodickou složku a také zda je či není řadou náhodných čísel

Děkuji za pozornost...