Statistické metody a zpracování dat

Analýza časových řad III.

Petr Dobrovolný

Autokorelace časových řad

Autokorelační analýza - metoda, kterou lze zkoumat vzájemné vztahy mezi hodnotami jedné časové řady.

Může sloužit jako metoda k definování sezónní a cyklické složky časových řad.

Jejím základem je výpočet **autokorelačního koeficientu**, resp. **autokorelační funkce**.

Autokorelační koeficient

Autokorelační koeficient r_k je 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**. Je to tedy korelační koeficient vypočtený mezi jednotlivými členy časové řady, mezi kterými je *k-1* jiných pozorování tedy *lag* = *k* a označujeme ho jako autokorelační koeficient k-tého řádu.

Pro *k = 0* je hodnota $r_0 = 1$ - je to vlastně hodnota korelačního koeficientu.

Základní pojmy

Rozptyl (variance) – míra variability (proměnlivosti) statistického znaku *x*

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Kovariance – absolutní míra vzájemné variability dvou statistických znaků *x; y*

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Korelace - relativní míra vzájemné variability dvou statistických znaků *x; y*

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Základní vztahy

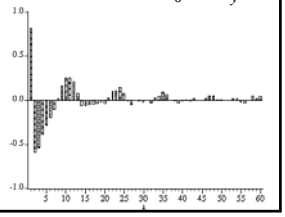
Autokovariance – absolutní míra proměnlivosti členů časové řady *y* posunutých o určitou hodnotu *k*.

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

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 – hodnoty $r_y(k)$ pro $k=1,2,\ldots M$, kde $M < N/2$, N – délka řady



Autokorelační funkce

Autokorelační funkce (ACF) je potom závislost mezi hodnotami autokorelačního koeficientu a hodnotami posunu k.

Vyjadřuje se formou grafu – tzv. **korelogramu** (viz. obrázek). Na ose x jsou hodnoty lag (k), na ose y hodnoty autokorelačního koeficientu.

Hodnoty autokorelační funkce se pohybují v intervalu –1,1.

ACF je vhodným nástrojem k posouzení, zda časová řada obsahuje cyklickou či periodickou složku a také zda je či není řadou náhodných čísel – tedy do jaké míry je možné ji extrapolovat (předpovídat).

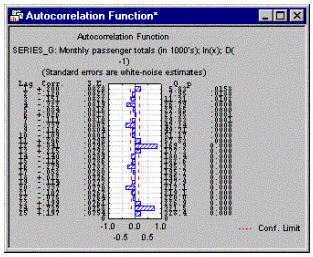
Interpretace ACF I

Korelogram bývá doplňován intervaly spolehlivosti, kterými lze hodnotit statistickou významnost autokorelačních koeficientů.

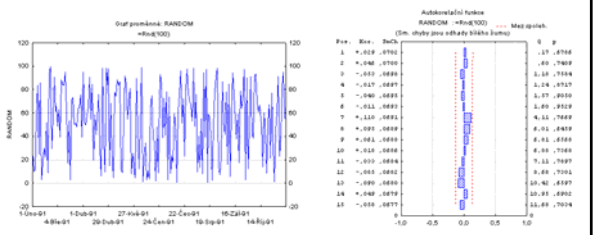
95 % interval spolehlivosti ACF lze z dostatečnou přesností zkonstruovat ze vztahu:

$$\pm 2/\sqrt{N}$$

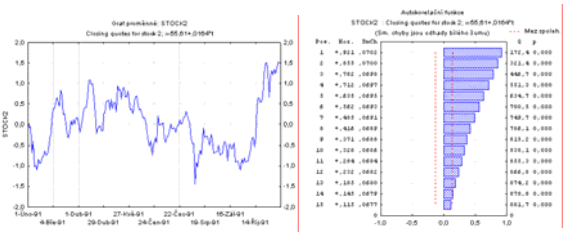
***N* – délka časové řady**



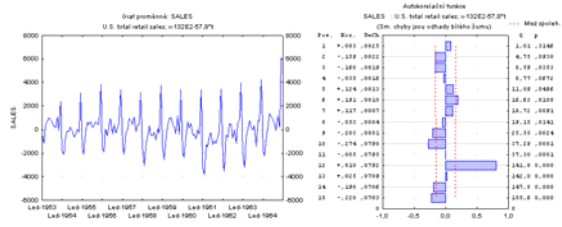
Časová řada náhodných čísel (bílý šum) a její autokorelační funkce



Časová řada bez periodické složky se silnou autokorelací a její autokorelační funkce

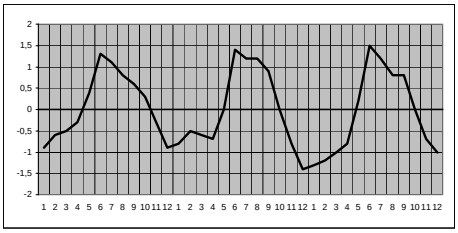


Časová řada obsahující výraznou sezónní složku a její autokorelační funkce



Spektrální analýza časových řad

Metody vycházejí z předpokladu, že řadu s výrazným periodickým kolísáním lze vyjádřit jako součet funkcí sin a cos s různou amplitudou a frekvencí.



Řadu na obrázku lze poměrně přesně aproximovat funkcí sin.

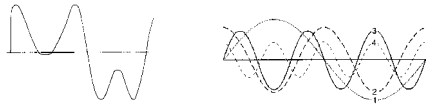
$$y_t = a \cdot \sin(\frac{2\pi}{T}t + \varphi)$$

Spektrální analýza časových řad

Reálné časové řady mívají složitější průběh. K jejich popisu lze použít **více členů** uvedeného obecného modelu s různou amplitudou a frekvencí.

Libovolný periodický pohyb s periodou T vzniká skládáním dvou či více harmonických pohybů, z nichž první má periodu T, další T/2, T/3 atd. Výsledkem je popis chování řady tzv. **Fourierovou řadou**.

$$y_t = a_1 \cdot \sin\frac{2\pi}{T}t + b_1 \cdot \cos\frac{2\pi}{T}t + a_2 \cdot \sin\frac{4\pi}{T}t + b_2 \cdot \cos\frac{4\pi}{T}t + a_3 \cdot \sin\frac{6\pi}{T}t + b_3 \cdot \cos\frac{6\pi}{T}t + \ldots$$



Řadu v levé části obrázku lze modelovat čtyřmi harmonickými pohyby uvedenými vpravo

Cíl spektrální analýzy

Cílem spektrální analýzy je získat obraz o **intenzitě zastoupení jednotlivých frekvencí** v časové řadě – o tzv. spektru řady.

Spektrum pojem převzatý z oblasti teorie vlnění. Paprsek „bílého“ světla tvoří náhodný - tzv. „**bílý šum**“. Ten lze rozložit na jednotlivé komponenty o různé amplitudě a frekvenci.

Spektrální analýza je takovým „hranolem“, kterým lze časovou řadu rozložit na jednotlivé komponenty. Na rozdíl od metod u kterých je délka cyklu (či spíše periody – sezónnosti) známá, spektrální analýza umožňuje **délku významných cyklů v řadě identifikovat**.

Základní pojmy

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.

$$f_i = \frac{1}{T_i}, i = 1,2,...,n \qquad y(t_i) \rightarrow y(f_i)$$

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

Rozklad časové řady na jednotlivé komponenty lze považovat za příklad lineární vícenásobné regrese.

Závisle proměnou představují členy časové řady a nezávisle proměnné představují *sin* a *cos* funkce všech jednotlivých frekvencí.

Ve shodě s výše uvedeným lze takovýto model lineární vícenásobné regrese vyjádřit následovně:

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

kde $f_k = k/q$

Princip spektrální analýzy

Analogicky jako v případě regresní závislosti parametry sin (a_k) a cos (b_k) funkcí jsou regresními koeficienty, které nám 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).

Jednoduchý příklad

Vytvoříme řadu o 16 členech:

$$y = 1.0*\cos(2\pi*0.0625*(t-1)) + 0.75*\sin(2\pi*0.2*(t-1))$$

pro $t = 1, 2, ...16$

Takto vytvořená řada obsahuje dvě periodické složky.

První má frekvenci a f=0,0626 - tzn. periodu 1/f = 16 – tedy celý cyklus trvá 16 časových jednotek).

Druhá periodická složka má frekvenci f = 0,2 (tj. periodu 5).

Koeficient funkce cos (1,0) je větší než koeficient funkce sin (0,75).

Jednoduchý příklad – pokračování

Výsledná tabulka spektrální analýzy**:**

Spectral analysis:VAR1 (shumex.sta)
No. of cases: 16

t	Freq-uency	Period	Cosine Coeffs	Sine Coeffs	Period-ogram
0	.0000		.000	0.000	.000
1	.0625	16.00	1.006	.028	8.095
2	.1250	8.00	.033	.079	.059
3	.1875	5.33	.374	.559	3.617
4	.2500	4.00	-.144	-.144	.333
5	.3125	3.20	-.089	-.060	.092
6	.3750	2.67	-.075	-.031	.053
7	.4375	2.29	-.070	-.014	.040
8	.5000	2.00	-.068	0.000	.037

Největší cos koeficient se vyskytuje na frekvenci 0,0625.

Menší sin koeficient na frekvenci 0,1875.

Tedy frekvence, které byly „vloženy“ do vytvořené řady se reflektují ve výstupní tabulce.

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).

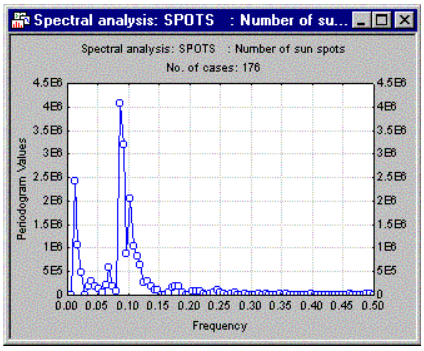
Periodogram

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

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

P_k – hodnota periodogramu na frekvenci f_k
N – počet členů časové řady

Hodnoty periodogramu mohou být interpretovány jako rozptyl (variance - suma čtverců) vstupních dat na dané frekvenci.



Hodnoty periodogramu jsou vykreslovány v grafu k příslušné frekvenci (periodě)

Problém „prosakování“ frekvencí (leakage)

V důsledku omezené délky zpracovávané řady se často stane, že periodogram vykazuje vysoké hodnoty na dvou blízkých frekvencích, které ve skutečnosti představují pouze jednu významnou sin či cos funkci na frekvenci, která padá mezi hodnoty vyjádřené v periodogramu.

V našem příkladě jsme do řady „vložili“ periodu o frekvenci 0,2, ve výsledku se však objevila vysoká hodnota koeficientu pro frekvenci 0,1875. Tento jev se označuje jako „leakage“ – „prosakování“ frekvencí.

Řešení problému „prosakování“ frekvencí

Padding (pad – vycpávka) – doplnění řady nulami. Protože hodnoty koeficientů se počítají pro frekvence určené jako N/t, lze bez vlivu na výsledky přidat na konec řady konstantu (nulu).

Např. ve výše uvedeném, příkladu přidáním deseti nul dostaneme největší hodnoty periodogramu právě pro „vložené“ frekvence 0,0625 a 0,2. Toto „prodloužení“ řady někdy ztěžuje interpretaci.

Tappering the series (taper – zužovat) - zkrácení délky řady na mocninu 2

Shlazení periodogramu a výpočet tzv. spektrální hustoty

Spektrální hustota

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.

Shlazení hodnot periodogramu

K nalezení nejvyšších spektrálních hustot v hodnotách periodogramu se využívá metod shlazení váženými klouzavými průměry.

Shlazovací okno má lichý počet (m) členů a existuje několik metod, které různým způsobem definují váhy.

Suma vah je rovna jedné a většina filtrů dává podobné výsledky.

Shlazení hodnot periodogramu

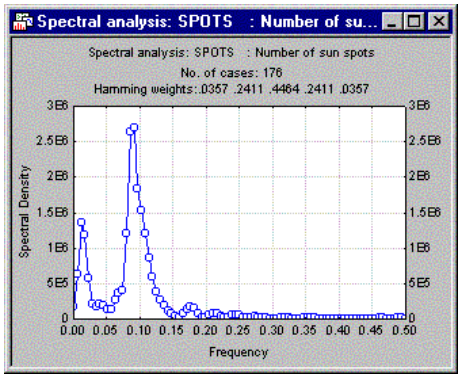
Příklad – určení vah tzv. Hammingova filtru pro okno s m členy, kdy p=(m-1)/2

$$w_j = 0{,}54 + 0{,}46 * \cos(2\pi * j / p) \qquad (j = 0,..., p)$$

$$w_{-j} = w_j$$

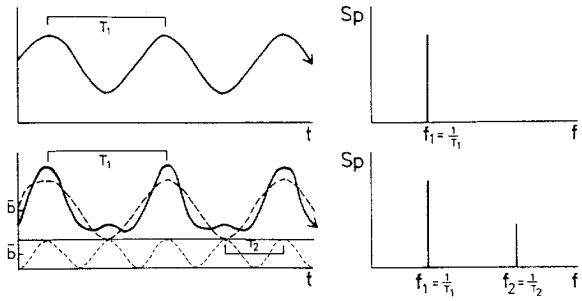
Váhy jsou symetrické a přiřazují nejvyšší hodnotu vždy střednímu členu shlazované části periodogramu.

Spektrální hustota

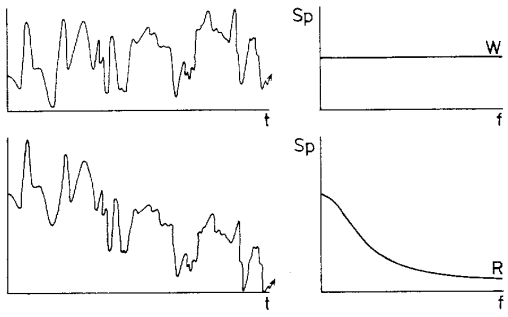


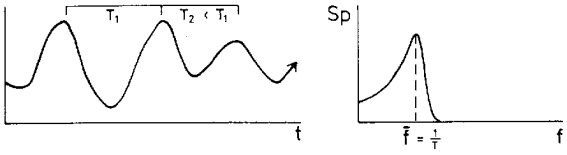
Příklady

Průběh časové funkce (vlevo) a spektrální funkce (vpravo) pro sinusoidu a dvě různé sinusoidy



Průběh časové funkce a spektrální funkce pro náhodná čísla a náhodná čísla (bílý šum) a náhodná čísla s trendem





Průběh časové funkce a spektrální funkce pro cyklické kolísání s proměnlivou délkou periody a amplitudy

Předzpracování časové řady

Odečtení průměru – pokud se průměr neodečte, bude vycházet výrazně vysoká hodnota periodogramu na frekvenci nula (0).

Odečtení trendu – jinak bude řada nestacionární

V některých případech je vhodné zvýraznit potenciální cykly shlazením časové řady metodou klouzavých průměrů.

Testování časové řady

Pokud řada neobsahuje žádný cyklus, znamená to, že hodnota každého členu řady je zcela nezávislá na hodnotách všech členů ostatních a řada představuje bílý šum.

Hodnoty periodogramu takovéto řady mají exponenciální rozdělení. Lze tedy provést test rozdělení hodnot periodogramu vůči exponenciálnímu rozdělení.

Lze také využít K-S testu (d testovací kritérium).