DIPLOMOVÁ PRÁCE

Matematicko–fyzikální fakulta Univerzity Karlovy

ANALÝZA HYDROMETEOROLOGICKÝCH ČASOVÝCH ŘAD

Zdeněk Hlávka

24. června 2011: překlad novou verzí LaTeXu
(čísla stránek se mohou lišit od původní verze z roku 1995)

OBOR : Matematika – Matematická statistika

VEDOUCÍ : Prof. RNDr. Jiří Anděl, DrSc.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a užil při tom pouze uvedenou literaturu.
Souhlasím se zapůjčováním této práce.

V Praze dne

. . . . . .

Rád bych poděkoval prof. RNDr. Jiřímu Andělovi, DrSc. za trpělivost, se kterou dohlížel na vznik této práce. Děkuji rovněž dr. Brůžkovi z Českého hydrometeorologického ústavu, který mi laskavě poskytl analyzovaná data.

Zdeněk Hlávka

Úvod

Cílem této diplomové práce je především provedení statistické analýzy hydrometeorologických časových řad. V teoretické části popíšu základní typy modelů časových řad a metody používané pro identifikaci modelu. Soustředím se zejména na autoregresní posloupnosti, které jsou poměrně dobře interpretovatelné a z výpočetního hlediska jednoduché.

V praktické části budu analyzovat tyto časové řady:

1. Roční průměry Wolfových čísel od roku 1749 do roku 1992
2. Průměrné roční teploty v Klementinu od roku 1771 do roku 1992
3. Roční úhrny srážek v Klementinu od roku 1876 do roku 1992

Mým cílem je nalezení vhodného autoregresního modelu, odhalení případné periodicity a otestování vzájemné závislosti těchto časových řad.

Všechny metody použité v praktické části popíšu v části teoretické.

Kapitola 1

Základní pojmy

Nechť $T \subseteq \mathbf{R}$. Potom náhodným procesem nazvu systém náhodných veličin $\{X_t, t \in T\}$. Nejčastěji se za množinu T bere množina celých čísel a mluví se o náhodné posloupnosti, nebo se za T bere množina reálných čísel a mluví se o spojitém náhodném procesu. Množina T se obvykle interpretuje jako čas.

Je-li $\{X_t\}$ komplexní náhodný proces, definuje se střední hodnota procesu jako funkce $\mu_t = EX_t$. Pokud platí, že $EX_t = 0$ pro všechna $t \in T$, řeknu, že proces $\{X_t\}$ je centrovaný.

Dále řeknu, že náhodný proces $\{X_t\}$ má konečné druhé momenty, platí-li $E|X_t|^2 < \infty$ pro všechna $t \in T$.

Má-li proces $\{X_t\}$ konečné druhé momenty, definuji kovarianční funkci náhodného procesu jako

$$R(s,t) = E(X_s - \mu_s)(\overline{X_t} - \overline{\mu_t}). \tag{1.1}$$

Hodnota $R(t,t)$ je rozptyl procesu v čase t.

Pokud kovarianční funkce $R(s,t)$ závisí na svých argumentech pouze prostřednictvím jejich rozdílu, zavádí se funkce jedné proměnné (která se také označuje R) vztahem

$$R(s-t) = R(s,t) \tag{1.2}$$

a o procesu $\{X_t\}$ říkáme, že je kovariančně stacionární. Je-li ještě navíc $\mu_t = \mu$ pro všechna $t \in T$, nazveme proces $\{X_t\}$ slabě stacionární. Většinou se slovo „slabě“ vynechává a hovoří se pouze o stacionárním procesu.

Rozptyl (slabě) stacionárního procesu je zřejmě $R(0)$. Je-li $R(0) \neq 0$, zavádí se u stacionárního procesu korelační funkce

$$B(t) = \frac{R(t)}{R(0)} \tag{1.3}$$

a u stacionární náhodné posloupnosti také parciální autokorelační funkce ϱ_{kk}, která se definuje jako parciální korelační koeficient X_t a X_{t+k} při pevných hodnotách $X_{t+1}, \ldots, X_{t+k-1}$.

Odhad autokorelační a parciální autokorelační funkce

Pro napozorovanou reálnou časovou řadu $x_1, \ldots, x_n$ se obvykle používá následující odhad střední hodnoty

$$\bar{x} = \sum_{t=1}^{n} \frac{x_t}{n}, \tag{1.4}$$

následující odhad autokovarianční funkce

$$c_k = \sum_{t=1}^{n-k} \frac{(x_t - \bar{x})(x_{t+k} - \bar{x})}{n}, \qquad \text{pro } k = 0, 1, \ldots, n-1 \tag{1.5}$$

a následující odhad autokorelační funkce

$$r_k = \frac{c_k}{c_0}, \qquad \text{pro } k = 0, 1, \ldots, n-1. \tag{1.6}$$

Pro výběrový korelační koeficient r^* počítaný z n nezávislých dvojic se stejným regulárním dvojrozměrným normálním rozdělením a s korelačním koeficientem $\varrho = 0$ platí, že $Er^* = 0$ a var $r^* = \frac{1}{n} + O(n^{-3/2})$ (viz Cramér [6]). Proto se také směrodatná odchylka $\sigma(r_k)$ odhadu r_k autokorelační funkce $B(k)$ většinou aproximuje pomocí

$$\hat{\sigma}(r_k) = \sqrt{\frac{1}{n}}. \tag{1.7}$$

Máme-li pak rozhodnout, zda $B(k) = 0$, porovnáme hodnotu $|r_k|$ obvykle s číslem $2\hat{\sigma}(r_k)$. Využitím asymptotické normality odhadu r_k tak získáme jednoduchý test na přibližně pěti-procentní hladině pravděpodobnosti.

Odhady r_{kk} parciální autokorelační funkce ϱ_{kk} se obvykle počítají rekurentně podle následujících vzorců (viz [5]):

$$r_{11} = r_1,$$

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j}, \qquad \text{pro } k > 1, \tag{1.8}$$

kde

$$r_{k,j} = r_{k-1,j} - r_{kk} r_{k-1,k-j}, \qquad \text{pro } j = 1, 2, \ldots, k-1.$$

Quenouille [19] navrhl aproximaci $\hat{\sigma}(r_{kk})$ pro směrodatnou odchylku odhadu r_{kk}. Je-li $\varrho_{kk} = 0$ pro $k > k_0$, pak

$$\hat{\sigma}(r_{kk}) = \sqrt{\frac{1}{n}}, \qquad \text{pro } k > k_0. \tag{1.9}$$

Nulovost ϱ_{kk} zamítáme, pokud hodnota $|r_{kk}|$ překročí dvojnásobek této směrodatné odchylky.

1.1 Příklady stacionárních posloupností

Posloupnost nekorelovaných náhodných veličin

Posloupnost nekorelovaných náhodných veličin $\{\varepsilon_t\}$ s nulovými středními hodnotami a stejným rozptylem $\sigma^2 = E|\varepsilon_t|^2$ je nejjednodušším příkladem stacionární posloupnosti. Její kovarianční funkce je zřejmě rovna $R(t) = \sigma^2$ pro $t = 0$, $R(t) = 0$ pro $t \neq 0$. Kvůli jejím spektrálním vlastnostem se posloupnosti $\{\varepsilon_t\}$ říká bílý šum.

Lineární proces

Nechť $\{\varepsilon_t\}$ je bílý šum z předchozího příkladu a $c_0, c_1, \ldots$ jsou taková čísla, že $\sum_{k=0}^{\infty} |c_k|^2 < \infty$. Lineární proces je definován jako

$$X_t = \sum_{k=0}^{\infty} c_k \varepsilon_{t-k}, \qquad \text{pro } t = \ldots, -1, 0, 1, \ldots. \tag{1.10}$$

Kovarianční funkce lineárního procesu je

$$R(t) = \sigma^2 \sum_{j=0}^{\infty} c_{t+j} \overline{c_j}, \qquad \text{pro } t = 0, 1, \ldots. \tag{1.11}$$

Pro $t < 0$ použijeme vztah $R(-t) = \overline{R(t)}$.

Posloupnost klouzavých součtů řádu n

Posloupnost klouzavých součtů je speciálním příkladem lineárního procesu. Nechť $\{\varepsilon_t\}$ je bílý šum a $a_0, a_1, \ldots, a_n$ jsou konstanty. Předpokládejme, že $a_0 \neq 0$, $a_n \neq 0$. Posloupnost klouzavých součtů řádu n je určena následujícím vztahem

$$X_t = \sum_{k=0}^{n} a_k \varepsilon_{t-k}, \qquad \text{pro } t = \ldots, -1, 0, 1, \ldots. \tag{1.12}$$

Zřejmě platí $EX_t = 0$. Pro kovarianční funkci platí

$$R(t) = \sigma^2 \sum_{j=0}^{n-t} a_{t+j} \overline{a_j}, \qquad \text{pro } t = 0, \ldots, n \tag{1.13}$$

Pro $|t| > n$ je $R(t) = 0$. Pro $t < 0$ použijeme vztah $R(-t) = \overline{R(t)}$.

Skutečnost, že kovarianční funkce je nulová pro $t > n$, se používá k identifikaci modelu klouzavých součtů. Je dobré si uvědomit, že v praxi máme většinou k dispozici pouze odhad kovarianční funkce, který přesně nulových hodnot nabývat nemusí. Většinou pracujeme s odhady autokorelační funkce, jejíž nulovost můžeme snadno otestovat pomocí aproximace (1.7).

Posloupnost klouzavých součtů řádu n se značí symbolem MA(n). Označení vzniklo z anglického „moving average“.

Kapitola 2

Testy náhodnosti

Ještě před tím, než začneme hledat pro napozorovanou časovou řadu vhodný model, je dobré otestovat, zda se nejedná o bílý šum. K tomuto účelu můžeme použít následující testy náhodnosti.

Test založený na bodech růstu

Tento test je založený na počtu bodů v nichž daná řada $X_1, \ldots, X_n$ roste. Pokud jsou některé sousední hodnoty stejné, tak je až na jednu z nich vyškrtneme. Definujme náhodné veličiny V_t pro $t = 1, 2, \ldots, n-1$ vztahem

$$\begin{aligned} V_t &= 1, \qquad \text{pokud } X_{t+1} > X_t, \\ V_t &= 0, \qquad \text{pokud } X_{t+1} < X_t. \end{aligned} \tag{2.1}$$

Za platnosti hypotézy náhodnosti zřejmě $P(V_t = 1) = P(V_t = 0) = 1/2$ a pro střední hodnotu Ek počtu bodů růstu $k = \sum_{t=1}^{n-1} V_t$ platí

$$Ek = E \sum_{t=1}^{n-1} V_t = \sum_{t=1}^{n-1} EV_t = \frac{n-1}{2}. \tag{2.2}$$

Jednoduše lze ukázat, že za platnosti nulové hypotézy $P(V_t \cdot V_{t+1} = 1) = 1/6$ a $P(V_t \cdot V_{t+s} = 1,\ s \geq 2) = 1/4$. Zřejmě $V_t^2 = V_t$. Pro rozptyl počtu bodů růstu k platí

$$\text{var } k = E \left(\sum_{t=1}^{n-1} V_t \right)^2 - \left(E \sum_{t=1}^{n-1} V_t \right)^2.$$

To můžeme rozepsat jako

$$\text{var } k = E \left(\sum_{t=1}^{n-1} V_t^2 + 2 \sum_{t=1}^{n-2} V_t V_{t+1} + \sum_{t=1}^{n-1} \sum_{j=1, j \neq t, t\pm 1}^{n-1} V_t V_j \right) - \left(E \sum_{t=1}^{n-1} V_t \right)^2.$$

Dále

$$\text{var } k = \sum_{t=1}^{n-1} EV_t^2 + 2 \sum_{t=1}^{n-2} EV_t V_{t+1} + \sum_{t=1}^{n-1} \sum_{j=1, j \neq t, t\pm 1}^{n-1} EV_t V_j - \left(E \sum_{t=1}^{n-1} V_t \right)^2.$$

Dosadíme za střední hodnoty

$$\text{var } k = (n-1)\frac{1}{2} + 2(n-1)\frac{1}{6} + [(n-3)(n-4) + 2(n-3)]\,\frac{1}{4} - \left(\frac{n-1}{2}\right)^2$$

a po úpravách dostaneme

$$\text{var } k = \frac{n+1}{12}. \tag{2.3}$$

Rozdělení veličiny k je tabelováno pro $n \leq 100$ v tabulkách [14]. Pro větší n můžeme využít asymptotickou normalitu k (viz Wolfowitz [28]) a náhodnost dat zamítáme na dané hladině pravděpodobnosti, pokud veličina

$$\frac{\left|k - \frac{n-1}{2}\right|}{\sqrt{\frac{n+1}{12}}} \tag{2.4}$$

překročí příslušnou kritickou hodnotu normálního rozdělení. Pokud tento test zamítá náhodnost dat, pak máme podezření zejména na existenci trendové složky. Pokud je bodů růstu příliš málo, znamená to, že se data s časem zmenšují, pokud je bodů růstu příliš mnoho, data se naopak s časem zvětšují. Tento test nemá příliš velkou sílu, pokud časová řada obsahuje periodickou složku, protože ta počet bodů růstu příliš neovlivní.

Test založený na bodech zvratu

Pokud jsou v dané řadě $X_1, \ldots, X_n$ některé sousední hodnoty stejné, tak je opět až na jednu z nich vyškrtneme. Říkáme, že X_t je bod zvratu, pokud $X_{t-1} < X_t > X_{t+1}$ nebo $X_{t-1} > X_t < X_{t+1}$.

Definujme náhodné veličiny U_t pro $t = 1, 2, \ldots, n-2$ jako

$$\begin{aligned} U_t &= 1, \qquad \text{pokud } X_t < X_{t+1} > X_{t+2} \text{ nebo } X_t > X_{t+1} < X_{t+2} \\ U_t &= 0 \qquad \text{jinak .} \end{aligned} \tag{2.5}$$

Počet všech bodů zvratu označme $r = \sum_{t=1}^{n-2} U_t$. Za platnosti hypotézy náhodnosti lze jednoduše ukázat, že $P(U_t = 1) = 2/3$, $P(U_t U_{t+1} = 1) = 5/12$, $P(U_t U_{t+2} = 1) = 27/60$ a $P(U_t U_{t+s} = 1, s \geq 3) = 4/9$. Střední hodnotu počtu bodů zvratu můžeme spočítat jako

$$Er = E\left(\sum_{t=1}^{n-2} U_t\right) = \sum_{t=1}^{n-2} EU_t = (n-2)\frac{2}{3}. \tag{2.6}$$

Rozptyl spočítáme zcela obdobně jako u bodů růstu

$$\text{var } r = E\left(\sum_{t=1}^{n-2} U_t\right)^2 - \left(E\sum_{t=1}^{n-2} U_t\right)^2.$$

To můžeme rozepsat jako

$$\text{var } r = E\left(\sum_{t=1}^{n-2} U_t^2 + 2\sum_{t=1}^{n-3} U_t U_{t+1} + 2\sum_{t=1}^{n-4} U_t U_{t+2} + \sum_{t=1}^{n-1}\sum_{j=1, j\neq t, t\pm 1, t\pm 2}^{n-1} U_t U_j\right) -$$

$$- \left(E \sum_{t=1}^{n-1} U_t \right)^2 .$$

Dále

$$\text{var } r = \sum_{t=1}^{n-2} EU_t + 2 \sum_{t=1}^{n-3} EU_t U_{t+1} + 2 \sum_{t=1}^{n-3} EU_t U_{t+1} + \sum_{t=1}^{n-2} \sum_{j=1, j\neq t, t\pm 1, t\pm 2}^{n-1} EU_t U_j -$$

$$- \left(E \sum_{t=1}^{n-2} U_t \right)^2 .$$

Dosadíme za střední hodnoty

$$\text{var } r = (n-2)\frac{2}{3} + 2(n-3)\frac{5}{12} + 2(n-4)\frac{27}{60} +$$

$$+ \left[(n-6)(n-7) + 2(n-6) + 2(n-5) \right] \frac{4}{9} - \left[\frac{2(n-2)}{3} \right]^2$$

a po úpravách dostaneme

$$\text{var } r = \frac{16n - 29}{90}. \tag{2.7}$$

Při větším n hypotézu náhodnosti zamítáme, pokud veličina

$$\frac{\left| r - \frac{2(n-2)}{3} \right|}{\sqrt{\frac{16n-29}{90}}} \tag{2.8}$$

překročí příslušnou kritickou hodnotu normálního rozdělení. Příliš malý počet bodů zvratu může být způsoben zejména rostoucí nebo klesající tendencí dat, velký počet bodů zvratu může být způsoben například přítomností periodické složky s velkou frekvencí.

Spearmanův koeficient pořadové korelace

Označme $q_1, \ldots, q_n$ pořadí hodnot řady $X_1, \ldots, X_n$. Číslo q_i udává pořadí čísla X_i v uspořádané řadě $X_{[1]} \leq \ldots \leq X_{[n]}$. Pokud jsou některá čísla X_i stejná, pak se pořadí uvnitř skupin stejných členů určuje náhodně.

Spearmanův koeficient pořadové korelace ϱ se definuje jako korelační koeficient posloupnosti $(1, q_1), \ldots, (n, q_n)$

$$\varrho = \frac{\sum_{i=1}^{n} \left(i - \frac{1}{n} \sum_{j=1}^{n} j \right) \left(q_i - \frac{1}{n} \sum_{j=1}^{n} q_j \right)}{\sqrt{\sum_{i=1}^{n} \left(i - \frac{1}{n} \sum_{j=1}^{n} j \right)^2 \sum_{i=1}^{n} \left(q_i - \frac{1}{n} \sum_{j=1}^{n} q_j \right)^2}}. \tag{2.9}$$

Tento vztah můžeme zjednodušit na

$$\varrho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (i - q_i)^2 \tag{2.10}$$

Pokud je předpoklad náhodnosti porušen vzestupnou tendencí čísel X_i, projeví se to kladnou korelovaností pořadí $q_1, \ldots, q_n$ a čísel $1, 2, \ldots, n$ a tedy většími hodnotami korelačního koeficientu ϱ, naopak v případě sestupné tendence se porušení předpokladu náhodnosti projeví malými hodnotami koeficientu ϱ.

Pro $n \leq 30$ jsou kritické hodnoty pro Spearmanův koeficient tabelovány v knize [2]. Při větším n zamítáme hypotézu náhodnosti, pokud veličina $|\varrho|\sqrt{n-1}$ překročí příslušnou kritickou hodnotu normálního rozdělení. Podrobně je tato metoda vyložena v knize [7].

Mediánový test — ověřování náhodnosti podle počtu sérií

Označme výběrový medián řady $X_1, \ldots, X_n$ jako M. Pozorování, která leží přímo na mediánu, z testované řady vyškrtneme. Sérií nazveme co největší skupinu sousedních pozorování, která leží všechna nad (respektive všechna pod) mediánem. Počet sérií označme jako u, celkový počet pozorování pod mediánem jako m_p a celkový počet pozorování nad mediánem jako m_n.

Pro m_p a m_n menší než 20 jsou odpovídající kritické hodnoty uvedené v tabulkách [12]. Pro větší m_p a m_n (pokud nejsou příliš různá) hypotézu náhodnosti zamítáme, pokud veličina

$$\frac{|u - \frac{2m_p m_n}{m_p+m_n} - 1|}{\sqrt{\frac{2m_p m_n (2m_p m_n - m_p - m_n)}{(m_p+m_n)^2 (m_p+m_n-1)}}} \tag{2.11}$$

překročí příslušnou kritickou hodnotu normálního rozdělení (viz Fabian [7]).

Příliš malý počet sérií vzniká zejména v případě vzestupné nebo sestupné tendence, nebo pokud je v řadě obsažena periodická složka s dlouhou periodou. Velký počet sérií může být způsoben například periodickou složkou s velkou frekvencí.

Test založený na odhadu rozptylu ze sousedních pozorování

Tímto testem můžeme testovat hypotézu, že normálně rozdělené náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé a stejně rozdělené. Test je založen na porovnání odhadu rozptylu

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \frac{1}{n} \sum_{j=1}^{n} X_j)^2 \tag{2.12}$$

s hodnotou

$$q = \frac{1}{n-1} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2. \tag{2.13}$$

Pokud je splněn předpoklad náhodnosti, budou s^2 a q nabývat zhruba stejných hodnot. Pokud jsou si sousední hodnoty příliš blízké (např. rostoucí tendence, klesající tendence,

dlouhá periodicita), bude q podstatně menší než s^2. Pokud jsou sousední hodnoty vzdálené víc, než by vyplývalo z náhodnosti dat, bude i q větší než s^2. K porovnání hodnot q a s^2 se používá statistika

$$d = \frac{n \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}. \tag{2.14}$$

Kritické hodnoty statistiky d jsou pro $n \leq 20$ tabelovány v [7]. Při větším n zamítáme nulovou hypotézu, pokud veličina

$$\frac{1}{2} \frac{|d-2|}{\sqrt{\frac{n-2}{n^2-1}}} \tag{2.15}$$

překročí příslušnou kritickou hodnotu normálního rozdělení (viz Fabian [7]).

Kapitola 3

Autoregresní posloupnosti

Nechť $\{\varepsilon_t\}$ je reálný bílý šum s rozptylem σ^2 a nechť $\varphi_1, \ldots, \varphi_p$ jsou taková reálná čísla, že polynom

$$P(z) = z^p - \varphi_1 z^{p-1} - \ldots - \varphi_p \tag{3.1}$$

má všechny kořeny uvnitř jednotkového kruhu[1]. Autoregresní posloupnost $\{X_t\}$ řádu p značená jako AR(p) je definovaná jako

$$X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \varepsilon_t. \tag{3.2}$$

Střední hodnota autoregresní posloupnosti AR(p) je zřejmě nulová a její autokorelační funkci $B(t)$ můžeme spočítat například tak, že obě strany vztahu (3.2) vynásobíme X_{t-k} a přejdeme ke středním hodnotám. Dostaneme

$$EX_tX_{t-k} = \varphi_1 EX_{t-1}X_{t-k} + \ldots + \varphi_p EX_{t-k}X_{t-p} + E\varepsilon_t X_{t-k}.$$

Protože zřejmě platí $E\varepsilon_t X_{t-k} = 0$ pro $k > 0$, můžeme psát

$$R(k) = \varphi_1 R(k-1) + \ldots + \varphi_p R(k-p) \qquad \text{pro } k = 1, 2, \ldots.$$

Pokud všechny tyto rovnice vydělíme $R(0)$, dostaneme tzv. Yule-Walkerovu soustavu rovnic

$$B(t) = \varphi_1 B(t-1) + \varphi_2 B(t-2) + \ldots + \varphi_p B(t-p), \qquad \text{pro } t = 1, 2, \ldots. \tag{3.3}$$

Protože platí, že $B(0) = 1$ a $B(t) = B(-t)$, dostaneme z (3.3) pro $t = 1, \ldots, p-1$ soustavu $p-1$ rovnic pro $p-1$ neznámých $B(1), \ldots, B(p-1)$. Pro $t \geq p$ je (3.3) homogenní lineární diferenční rovnice, kterou řešíme pomocí obecné teorie diferenčních rovnic.

Rozptyl $R(0)$ autoregresní posloupnosti AR(p) můžeme spočítat podobně. Vztah (3.2) vynásobíme X_t a přejdeme ke středním hodnotám

$$EX_t^2 = \varphi_1 EX_{t-1}X_t + \ldots + \varphi_p EX_{t-p}X_t + E\varepsilon_t X_t.$$

[1] Podmínka stacionarity pro autoregresní posloupnost.

Tento vztah můžeme dále upravit jako

$$R(0) = \varphi_1 R(1) + \ldots + \varphi_p R(p) + E\varepsilon_t X_t$$

a využitím vztahu $R(k) = R(0)B(k)$ dostaneme

$$R(0) = \varphi_1 R(0)B(1) + \ldots + \varphi_p R(0)B(p) + E\varepsilon_t X_t,$$

takže

$$R(0) = \frac{E\varepsilon_t X_t}{1 - \varphi_1 B(1) - \ldots - \varphi_p B(p)}. \tag{3.4}$$

Znásobíme-li (3.2) ε_t a přejdeme ke středním hodnotám, dostaneme $E\varepsilon_t X_t = E\varepsilon_t^2 = \sigma^2$ a pro $R(0)$ máme vzorec

$$R(0) = \frac{\sigma^2}{1 - \varphi_1 B(1) - \ldots - \varphi_p B(p)}. \tag{3.5}$$

Velmi snadno pak můžeme spočítat i autokovarianční funkci, protože

$$R(t) = R(0)B(t). \tag{3.6}$$

3.1 Odhad parametrů a odlehlá pozorování

Uvažujme autoregresní posloupnost řádu p

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + \varepsilon_t, \tag{3.7}$$

kde $\{\varepsilon_t\}$ je normální bílý šum s rozptylem σ^2. Označme

$$\mathbf{z}_t^T = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})$$

a

$$\mathbf{\Phi}^T = (\varphi_1, \varphi_2, \ldots, \varphi_p).$$

Teď můžeme vztah (3.7) přepsat tímto způsobem

$$X_t = \mathbf{z}_t^T \mathbf{\Phi} + \varepsilon_t. \tag{3.8}$$

Máme-li pozorování $X_1, \ldots, X_n$, dostaneme následující soustavu rovnic

$$\mathbf{X} = \mathbf{\Gamma\Phi} + \boldsymbol{\varepsilon}, \tag{3.9}$$

kde $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ je vektor pozorování, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ je chybový vektor a

$$\mathbf{\Gamma} = \begin{pmatrix} X_0 & X_{-1} & \ldots & X_{-p+1} \\ X_1 & X_0 & \ldots & X_{-p+2} \\ \vdots & \vdots & \ldots & \vdots \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{pmatrix} \tag{3.10}$$

je tzv. matice experimentu. Hodnoty $X_i, i \leq 0$ se volí pevně, nejjednodušší možnost je jejich střední hodnota, tj. nula[2].

Odhad $\hat{\mathbf{\Phi}}$ vektoru parametrů $\mathbf{\Phi}$ metodou nejmenších čtverců je určen vztahem

$$\hat{\mathbf{\Phi}} = (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}. \tag{3.11}$$

Nechť $\hat{\mathbf{X}} = \mathbf{\Gamma}\hat{\mathbf{\Phi}}$ jsou vyrovnané hodnoty. Vektor residuí $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n)^T$, je definován jako

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{X} - \hat{\mathbf{X}}. \tag{3.12}$$

Dále platí $\hat{\mathbf{X}} = \mathbf{\Gamma}\hat{\mathbf{\Phi}} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}$, neboli

$$\hat{\mathbf{X}} = \mathbf{H}\mathbf{X}, \tag{3.13}$$

kde $\mathbf{H} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T$. Matici $\mathbf{H}$ se v anglické literatuře říká „hat matrix“, protože „pokládá stříšku na $\mathbf{X}$“. Česky se matici $\mathbf{H}$ nejčastěji říká „projekční“.

Vektor residuí $\hat{\boldsymbol{\varepsilon}}$ může být také jednoduše vyjádřen pomocí projekční matice $\mathbf{H}$. Platí totiž vztahy

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= (\mathbf{I} - \mathbf{H})\mathbf{X}, \\ \hat{\boldsymbol{\varepsilon}} &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}. \end{aligned} \tag{3.14}$$

Jako odhad rozptylu bílého šumu σ^2 se používá

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n}. \tag{3.15}$$

Od této chvíle budu značit diagonální prvky matice $\mathbf{H}$ jako h_t, tedy

$$h_t = \mathbf{z}_t^T(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{z}_t. \tag{3.16}$$

Dále platí, že pokud $n \to \infty$, pak $n^{-1}(\mathbf{\Gamma}^T\mathbf{\Gamma}) \to \mathbf{\Sigma}$ v pravděpodobnosti (viz Tong [25]), kde

$$\mathbf{\Sigma} = \begin{pmatrix} R(0) & R(1) & \ldots & R(p-1) \\ R(1) & R(0) & \ldots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \ldots & R(0) \end{pmatrix}$$

je kovarianční matice. Definujeme-li $d_t = \mathbf{z}_t^T\mathbf{\Sigma}^{-1}\mathbf{z}_t$ pro $t = 1, 2, \ldots, n$, pak

$$nh_t = \mathbf{z}_t^T\left(\frac{\mathbf{\Gamma}^T\mathbf{\Gamma}}{n}\right)^{-1}\mathbf{z}_t \tag{3.17}$$

konverguje v pravděpodobnosti k d_t. To znamená, že nh_t můžeme interpretovat jako jistou vzdálenost mezi $\mathbf{z}_t$ a nulovým vektorem (tj. vektorem středních hodnot).

[2]Někdy se k volbě $X_i, i \leq 0$ využívá tzv. princip zpětné extrapolace (viz [1],[5]).

Přístup k detekci odlehlých pozorování, založený na diagonálních prvcích projekční matice

Následující postup navrhl Tong v knize [25]. Autoregresní posloupnost (3.7) můžeme zapsat v následujícím tvaru

$$\mathbf{z}_{t+1} = \mathbf{B}\mathbf{z}_t + \tilde{\boldsymbol{\varepsilon}}_t \tag{3.18}$$
$$X_t = (1, 0, \ldots, 0)\mathbf{z}_{t+1},$$

kde $\tilde{\boldsymbol{\varepsilon}}_t = (\varepsilon_t, 0, \ldots, 0)^T$ a

$$\mathbf{B} = \begin{pmatrix} \varphi_1 & \varphi_2 & \ldots & \varphi_{p-1} & \varphi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}.$$

Nyní předpokládejme, že máme realizaci $X_1, X_2, \ldots, X_n$ autoregresní posloupnosti AR(p). Z hlediska zápisu (3.18) je mnohem rozumnější zabývat se polohou vektoru $\mathbf{z}_{t+1} = (X_t, X_{t-1}, \ldots, X_{t-p+1})$ v p-rozměrném prostoru, než jenom velikostí jeho první souřadnice X_t. Proto bychom při detekci odlehlých pozorování v autoregresní posloupnosti měli hledat odlehlý vektor $\mathbf{z}_t$.

K měření „odlehlosti“ $\mathbf{z}_t$ můžeme použít vzdálenost d_t, kde

$$d_t = \mathbf{z}_t^T \boldsymbol{\Sigma}^{-1} \mathbf{z}_t. \tag{3.19}$$

Za platnosti hypotézy, že autoregresní posloupnost je gaussovská a že se v ní nevyskytuje žádné odlehlé pozorování, má vektor $\mathbf{z}_t$ p-rozměrné normální rozdělení s varianční maticí $\boldsymbol{\Sigma}$, a tedy d_t má rozdělení χ^2_p pro $t = p+1, \ldots, n$. Proto, díky (3.17), můžeme pro větší n použít nh_t jako užitečný nástroj pro detekci odlehlých pozorování. Pokud je nh_t větší než příslušná kritická hodnota rozdělení χ^2_p, můžeme říct, že $\mathbf{z}_t = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})^T$ je odlehlý vektor.

Předpokládejme, že náhodná veličina X_{t-1} je velká; pak ovlivní vektory $\mathbf{z}_t, \mathbf{z}_{t+1}, \ldots, \mathbf{z}_{t+p-1}$ a čísla $h_t, h_{t+1}, \ldots, h_{t+p-1}$ budou velká. Proto, jestliže h_{t-1} je malé a h_t je velké, můžeme identifikovat X_{t-1} jako odlehlé pozorování. Bohužel při tomto postupu dochází ke kumulaci obyčejných testů a tak jeho výsledky musíme interpretovat opatrně.

3.2 Určení řádu autoregresního modelu

Prvním problémem, na který se narazí při praktickém hledání vhodného autoregresního modelu, je otázka určení správného řádu autoregrese.

Přístup založený na parciální autokorelační funkci

Tento přístup je založen na faktu, že parciální autokorelační funkce ϱ_{kk} autoregresní posloupnosti řádu p, AR(p), je nulová pro všechna $k > p$. Hledáme tzv. identifikační bod k_0, tj.

takový bod pro který platí $\varrho_{kk} = 0$ pro $k > k_0$. V praxi pracujeme pouze s odhadem parciální autokorelační funkce r_{kk}, který samozřejmě přesně nulových hodnot nabývat nemusí. Nulovost r_{kk} můžeme testovat pomocí Quenouilleovy aproximace pro směrodatnou odchylku r_{kk}, viz (1.9). Pokud se nám podaří najít rozumný identifikační bod k_0, volíme řád autoregrese právě jako k_0. Je však nutné mít na paměti, že pracujeme pouze s odhadnutými hodnotami parciální autokorelační funkce, takže naše závěry mohou být zkreslené.

Můžeme se také podívat na odhad r_k autokorelační funkce $B(k)$. Pokud se daná časová řada řídí autoregresním modelem, nemá autokorelační funkce identifikační bod. Významnost hodnot r_k můžeme posoudit pomocí aproximace (1.7). Podrobně je tato metoda vyložena v knize [5].

Přístup založený na odhadu rozptylu bílého šumu

Označme $\hat{\sigma}_k^2$ odhad (3.15) rozptylu bílého šumu za předpokladu, že uvažovaná řada se řídí modelem AR(k). Odhad $\hat{\sigma}_k^2$ má sestupnou tendenci a při dosažení správného řádu modelu začíná kolísat kolem správné hodnoty σ^2. Díky tomu můžeme aspoň přibližně odhadnout řád autoregrese. V praxi se neosvědčilo volit řád, pro který je odhad rozptylu bílého šumu minimální, protože při tomto postupu jsou neúměrně preferovány jako odhady řádu velké hodnoty k. Proto byly navrženy metody, které velké řády autoregrese penalizují.

Kritérium AIC

Toto kritérium pro hledání nejvhodnějšího modelu v určité předem zvolené třídě modelů odvodil Akaike na základě teorie informace ve tvaru

$$\text{AIC} = -2\ln(\text{maximalizovaná věrohodnost}) + 2(\text{počet parametrů}) \tag{3.20}$$

V případě hledání správného řádu autoregresního modelu (za předpokladu normality) se AIC používá ve tvaru (viz [24])

$$\text{AIC}(k) = \log \hat{\sigma}_k^2 + \frac{2k}{n}. \tag{3.21}$$

Odhad řádu modelu získaný minimalizací tohoto kritéria se označuje MAICE (z anglického Minimum AIC Estimator).

Kritérium BIC

Toto kritérium odvodili nezávisle na sobě Schwarz [22] na základě bayesovského přístupu a Rissanen [20] na základě informačního přístupu, který souvisí s hospodárným uložením potřebných informací o daném modelu s odhadovaným řádem v paměti samočinného počítače. BIC kritérium má tvar

$$\text{BIC}(k) = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}. \tag{3.22}$$

Název BIC je zkratka z anglického Bayesian Information Criterion.

Kritérium HQ

Kritérium HQ navrhli Hannan a Quinn [10], podle jejichž iniciál je označeno.

$$\mathrm{HQ}(k) = \log \hat{\sigma}_k^2 + ck\frac{\log(\log n)}{n} \tag{3.23}$$

Toto kritérium je založeno na zákonu iterovaného logaritmu. Hannan a Quinn ho používali pro $c = 1$ a dokázali jeho silnou konzistenci pro $c > 1$.

3.3 Ověřování modelu

Po zkonstruování autoregresního modelu je vhodné ověřit jeho platnost. K tomuto účelu se většinou používají metody založené na autokorelacích residuí $\hat{\varepsilon}_t$ ze vzorce (3.12).

Asi nejjednodušeji se správnost zkonstruovaného autoregresního modelu ověřuje pomocí tzv. portmanteau testu. „Portmanteau“ znamená francouzsky „věšák“ a toto označení prý vzniklo z toho, že na portmanteau statistiku Q jsou „navěšeny“ odhady autokorelační funkce residuí $r_k(\hat{\varepsilon})$. Portmanteau statistika se definuje jako

$$Q = n\sum_{k=1}^{K} r_k^2(\hat{\varepsilon}), \tag{3.24}$$

Číslo K se doporučuje volit tak, aby bylo srovnatelné s $\sqrt{n}$, kde n je délka analyzované řady. Velikost portmanteau statistiky je určena autokorelačními vlastnostmi residuí. Pokud jsme zvolili špatný model, tak residua nebudou navzájem nezávislá a jejich závislost se projeví v portmanteau statistice Q. Jestliže ověřujeme pomocí Q správný model AR(p), pak statistika Q má při větším n přibližně rozdělení χ^2_{K-p} (viz [5]).

Jako modifikace portmanteau statistiky byla navržena Box-Pierceova statistika ve tvaru

$$Q^* = n(n+2)\sum_{k=1}^{K} \frac{r_k^2(\hat{\varepsilon})}{n-k}. \tag{3.25}$$

Statistika Q^* má za platnosti ověřovaného modelu, stejně jako portmanteau statistika Q, asymptoticky rozdělení χ^2_{K-p} (viz např. [15]).

McLeod a Li v práci [16] navrhli pro test linearity podobný test založený na odhadnutých autokorelacích čtverců residuí

$$r_k^*(\hat{\varepsilon}) = \frac{\sum\limits_{j=1}^{n-k}(\hat{\varepsilon}_j^2 - \hat{\sigma}^2)(\hat{\varepsilon}_{j+k}^2 - \hat{\sigma}^2)}{\sum\limits_{j=1}^{n}(\hat{\varepsilon}_j^2 - \hat{\sigma}^2)^2}, \tag{3.26}$$

kde

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}{n}.$$

McLeod-Liova statistika

$$Q^{**} = n(n+2)\sum_{k=1}^{K}\frac{r_k^{*2}(\hat{\varepsilon})}{n-k} \tag{3.27}$$

má za platnosti ověřovaného modelu asymptoticky rozdělení χ^2_K.

Kapitola 4

Spektrální analýza

Při spektrální analýze se na časovou řadu díváme jako na nekonečnou směs periodických složek. Základní nástoje spektrální analýzy jsou spektrální hustota a periodogram, které udávají intenzitu zastoupení jednotlivých periodických složek v dané časové řadě. V celé této kapitole budu předpokládat, že $\{X_t\}$ je centrovaný (slabě) stacionární náhodný proces s konečnými druhými momenty a kovarianční funkcí $R(t)$.

Frekvence

Při analýze časových řad se frekvence (počet cyklů uskutečněných za časovou jednotku) udává v radiánech. Např. frekvence $\lambda = 2\pi$ radiánů znamená, že za časovou jednotku proběhne právě jeden cyklus. Odpovídající délku periody d (čas, za který proběhne jeden cyklus) dostaneme z jednoduchého vztahu $d = 2\pi/\lambda$. Další důležitý pojem je Nyquistova frekvence, tj. největší frekvence, kterou můžeme z pozorování konaných v diskrétních ekvidistantních časových okamžicích rozlišit. Nyquistova frekvence má hodnotu π, odpovídající délka periody je tedy 2. Pokud časová řada obsahuje periodické složky s frekvencí větší, než je Nyquistova frekvence, může se stát, že je vůbec neobjevíme (pokud je frekvence násobkem 2π), nebo že se budou projevovat jako frekvence menší (tomu se říká nerozlišitelnost frekvencí, anglicky aliasing).

4.1 Teorie

V této kapitole budu pracovat s komplexními náhodnými procesy.

Spektrální rozklad kovarianční funkce

Je-li $\{X_t\}$ komplexní stacionární posloupnost s kovarianční funkcí $R(t)$, pak existuje právě jedna neklesající, zleva spojitá funkce $F(\lambda)$, $F(-\pi) = 0$, $F(\pi) = R(0)$ taková, že

$$R(t) = \int_{-\pi}^{\pi} e^{it\lambda}\, dF(\lambda). \tag{4.1}$$

Toto vyjádření $R(t)$ se nazývá spektrální rozklad kovarianční funkce. Funkci $F(\lambda)$ se říká spektrální distribuční funkce. Pokud je $F(\lambda)$ absolutně spojitá funkce, pak existuje spektrální hustota $f(\lambda)$, která splňuje vztah

$$dF(\lambda) = f(\lambda)\ d\lambda$$

a spektrální rozklad (4.1) můžeme přepsat ve tvaru

$$R(t) = \int_{-\pi}^{\pi} e^{it\lambda} f(\lambda)\ d\lambda. \tag{4.2}$$

Spektrální hustota je asi nejdůležitější pojem spektrální analýzy, protože přímo udává intenzitu, s jakou je příslušná periodická složka zastoupena v časové řadě.

Náhodná míra a náhodný integrál

Mějme prostor $(\mathbf{R}, \mathbf{B})$, μ nechť je konečná míra na $\mathbf{B}$. Náhodná míra je definovaná jako množinová funkce Z na $\mathbf{B}$ taková, že pro $B \in \mathbf{B}$ je $Z(B)$ náhodná veličina splňující vlastnosti

1. $EZ(B) = 0$ pro všechna $B \in \mathbf{B}$
2. $B_1, B_2 \in \mathbf{B}, B_1 \bigcap B_2 = 0,$ pak $Z(B_1 \bigcup B_2) = Z(B_1) + Z(B_2)$ s.j.
3. $B_1, B_2 \in \mathbf{B},$ pak $EZ(B_1)Z(\bar{B}_2) = \mu(B_1 \bigcap B_2)$.

Nyní budu definovat náhodný integrál, nebo přesněji integrál funkce f podle náhodné míry Z: $\int f(\lambda)\ dZ(\lambda)$. Je-li $f(\lambda)$ jednoduchá měřitelná funkce, tj.

$$f(\lambda) = \sum_{k=1}^{n} a_k \chi_{I_k}(\lambda), \tag{4.3}$$

kde $a_1, \ldots, a_n$ jsou konstanty, $I_1, \ldots, I_n$ disjunktní borelovské množiny a χ_{I_k} označuje indikátor množiny I_k, pak

$$\int f(\lambda)\ dZ(\lambda) = \sum_{k=1}^{n} a_k Z(I_k). \tag{4.4}$$

Dále nechť $f(\lambda)$ je libovolná funkce z $L_2(\mu)$. Pak existuje posloupnost jednoduchých měřitelných funkcí $\{f_n\}$, která konverguje k f podle středu vzhledem k míře μ. Potom posloupnost náhodných veličin $\{\int f_n(\lambda)\ dZ(\lambda)\}$ konverguje podle středu a můžeme definovat

$$\int f(\lambda)\ dZ(\lambda) = l.i.m. \int f_n(\lambda)\ dZ(\lambda). \tag{4.5}$$

Spektrální rozklad stacionární posloupnosti

Centrovanou stacionární posloupnost $\{X_t\}$ se spektrální distribuční funkcí $F(\lambda)$ můžeme vyjádřit ve tvaru

$$X_t = \int_{-\pi}^{\pi} e^{it\lambda}\, dZ(\lambda) \qquad \text{pro } t = \ldots, -1, 0, 1, \ldots, \tag{4.6}$$

kde pro všechny borelovské podmnožiny B_1, B_2 intervalu $< -\pi, \pi >$ platí

$$EZ(B_1)Z(\bar{B}_2) = \int_{B_1 \bigcap B_2} dF(\lambda). \tag{4.7}$$

Toto vyjádření dostaneme ze spektrálního rozkladu kovarianční funkce (4.1) použitím Karhunenovy věty (viz Anděl [1]).

4.2 Periodogram a testování periodicity

Periodogram $I(\lambda)$ reálné časové řady $X_1, \ldots, X_n$ se definuje vzorcem

$$I(\lambda) = \frac{1}{2\pi n}[a^2(\lambda) + b^2(\lambda)], \qquad -\pi \le \lambda \le \pi, \tag{4.8}$$

kde

$$a(\lambda) = \sum_{t=1}^{n} X_t \cos(\lambda t)$$

$$b(\lambda) = \sum_{t=1}^{n} X_t \sin(\lambda t).$$

Při numerickém výpočtu se tento vzorec používá ve tvaru

$$I(\lambda) = \frac{1}{2\pi}\left(c_0 + 2\sum_{k=1}^{n-1} c_k \cos(k\lambda)\right), \tag{4.9}$$

kde $c_0, \ldots, c_{n-1}$ jsou odhady autokovarianční funkce podle vzorce (1.5).

Periodogram byl původně navržen jako vhodný nástroj pro hledání neznámých frekvencí $\lambda_1, \ldots, \lambda_p$ při konstrukci modelů typu

$$X_t = \sum_{j=1}^{p} [\alpha_j \cos(\lambda_j t) + \beta_j \sin(\lambda_j t)] + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{4.10}$$

Lze ukázat, že pro řadu X_t tohoto tvaru jsou hodnoty periodogramu malé a pohybují se přibližně kolem hodnoty $\sigma_\varepsilon^2/2\pi$, vyjma bodů $\lambda_1, \ldots, \lambda_p$, v nichž periodogram nabývá velkých hodnot řádu n (viz [1]).

Fisherův test

Pomocí Fisherova testu můžeme rozhodnout, jestli jsou některé hodnoty periodogramu významně velké. Nulová hypotéza má tvar

$$X_t = \varepsilon_t, \tag{4.11}$$

kde ε_t je normální bílý šum s rozptylem σ_ε^2 a testujeme ji proti alternativě

$$X_t = \sum_{j=1}^{p} [\alpha_j \cos(\lambda_j t) + \beta_j \sin(\lambda_j t)] + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{4.12}$$

kde $\sum_{j=1}^{p} (\alpha_j^2 + \beta_j^2) > 0$ a ε_t je normální bílý šum. s rozptylem σ_ε^2. Spočítáme hodnoty periodogramu v bodech

$$\lambda_j^* = \frac{2\pi j}{n} \qquad \text{pro } j = 1, \ldots, m, \tag{4.13}$$

kde m je největší celé číslo nepřesahující $(n-1)/2$, a normujeme je do tvaru

$$Y_j = \frac{I(\lambda_j^*)}{\sum\limits_{i=1}^{m} I(\lambda_i^*)} \qquad \text{pro } j = 1, \ldots, m. \tag{4.14}$$

Testová statistika má tvar

$$W = \max_{j=1,\ldots,m} Y_j. \tag{4.15}$$

Rozdělení statistiky W za platnosti nulové hypotézy je dáno výrazem

$$P(W > x) = m(1-x)^{m-1} - \binom{m}{2}(1-2x)^{m-1} + \ldots, \tag{4.16}$$

kde $0 < x < 1$ a na pravé straně sčítáme tak dlouho, dokud jsou členy $(1-kx)$ kladné. Důkaz lze nalézt např. v knize [1]. Pro $m \leq 50$ se doporučuje aproximace

$$P(W > x) = m(1-x)^{m-1}, \tag{4.17}$$

pro velká m můžeme použít aproximaci

$$P\left(Y > \frac{z + \ln m}{m}\right) = 1 - \exp\{-\exp(-z)\}. \tag{4.18}$$

Kritické hodnoty $f(m, \alpha)$ založené na těchto aproximacích mají pro zvolenou hladinu testu α tvar

$$f(m, \alpha) = 1 - \left(\frac{\alpha}{m}\right)^{1/(m-1)} \qquad \text{pro } m \leq 50 \tag{4.19}$$

a

$$f(m, \alpha) = \frac{\{-\ln[-\ln(1-\alpha)] + \ln m\}}{m} \qquad \text{pro } m \geq 50. \tag{4.20}$$

Nulovou hypotézu zamítáme, pokud statistika W překročí příslušnou kritickou hodnotu $f(m, \alpha)$. Jestliže chceme pokračovat v hledání dalších periodicit, můžeme použít postup, který navrhl Whittle [27]. Největší hodnotu periodogramu vynecháme a celý postup zopakujeme pro $m-1$. Takhle můžeme pokračovat tak dlouho, dokud nezjistíme všechny periodicity.

V praxi se ukázalo, že Fisherův test nemá příliš velkou sílu, pokud se v časové řadě vyskytuje více než jedna periodicita. Proto Siegel navrhl test, který tento nedostatek odstraňuje.

Siegelův test

Siegelův test je založen na statistice

$$T_\lambda = \sum_{j=1}^{m} [Y_j - \lambda f(m,\alpha)]^+ , \tag{4.21}$$

kde $(z)^+$ je kladná část čísla z a $0 < \lambda < 1$ je předem zvolená konstanta. Doporučuje se volit $\lambda = 0.6$. Rozdělení statistiky T_λ za platnosti nulové hypotézy je dáno vzorcem

$$P(T_\lambda > t) = \sum_{l=1}^{n} \sum_{k=0}^{l-1} (-1)^{k+l+1} \binom{n}{l} \binom{l-1}{k} \binom{n-1}{k} t^k \{[1 - \lambda l f(m,\alpha) - t]^+\}^{n-k-1}. \tag{4.22}$$

Tento vzorec odvodil Siegel [23].

4.3 Odhad spektrální hustoty

Za předpokladu spojitosti spektrální hustoty je jejím asymptoticky nestranným odhadem periodogram, bohužel však rozptyl periodogramu nekonverguje k nule a tak se periodogram může i pro velká n od správné spektrální hustoty značně lišit. Tuto nepříjemnou vlastnost periodogramu můžeme obejít použitím „vyrovnaného“ periodogramu tvaru

$$\hat{f}(\lambda_0) = \int_{-\pi}^{\pi} s(\lambda - \lambda_0) I(\lambda)\, d\lambda, \tag{4.23}$$

kde $s(\lambda)$ je omezená, po částech hladká funkce. Je-li $\{X_t\}$ normální reálná centrovaná stacionární posloupnost se spektrální hustotou $f(\lambda)$, pak je odhad (4.23) asymptoticky nestranný a asymptoticky konzistentní odhad výrazu

$$\int_{-\pi}^{\pi} s(\lambda - \lambda_0) f(\lambda)\, d\lambda, \tag{4.24}$$

ale volíme-li $s(\lambda)$ tak, aby byla soustředěna kolem bodu 0, můžeme dosáhnout toho, že tento výraz je dostatečně blízký odhadované hodnotě $f(\lambda_0)$. Je možné ukázat, že (4.23) lze ekvivalentně přepsat jako

$$\hat{f}(\lambda_0) = w_0 c_0 + 2 \sum_{k=1}^{n-1} w_k c_k \cos(\lambda_0 k), \tag{4.25}$$

kde c_k jsou odhadnuté autokovariance (1.5) a w_k jsou vhodně volené váhy. Často se používají Parzenovy váhy

$$\begin{aligned}
w_k &= \tfrac{1}{2\pi}\left(1 - \tfrac{6k^2}{K^2}\left(1 - \tfrac{k}{K}\right)\right) && \text{pro } k = 0, 1, \ldots, \tfrac{K}{2} \\
w_k &= \tfrac{1}{\pi}\left(1 - \tfrac{k}{K}\right)^3 && \text{pro } k = \tfrac{K}{2} + 1, \ldots, K \\
w_k &= 0 && \text{pro } k > K,
\end{aligned}$$

kde K se obvykle volí jako sudé číslo v rozmezí od $n/6$ do $n/5$.

Kapitola 5

Testování nezávislosti mezi časovými řadami

Často je zapotřebí testovat nezávislost dvou časových řad. Nejjednodušší testy jsou založené na korelačním koeficientu. V celé této kapitole budeme pracovat pouze se stacionárními časovými řadami.

5.1 Metody založené na korelaci

Vzájemná korelační funkce

Při vyšetřování nezávislosti dvou stacionárních řad $\{Y_t\}$ a $\{Z_t\}$ můžeme spočítat korelační koeficient r_0 z dvojic $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. Tento postup je však nevhodný, pokud se závislost mezi časovými řadami projevuje s určitým zpožděním. Proto se definuje výběrová vzájemná korelační funkce r_s jako výběrový korelační koeficient z časových řad $\{Y_{s+t}\}$ a $\{Z_t\}$ s nulovými středními hodnotami vzorcem

$$r_s = \frac{\sum_{k=\max(1,1-s)}^{\min(n,n-s)} Y_{s+k} Z_k}{\sqrt{\sum_{k=1}^{n} Y_k^2 \sum_{k=1}^{n} Z_k^2}} \qquad \text{pro } s = 0, \pm 1, \pm 2, \ldots, \pm(n-1). \tag{5.1}$$

Přestože až na výjimky není znám objektivní test pro vyhodnocení výběrové vzájemné korelační funkce r_s, můžeme z jejího průběhu někdy vyčíst důležitou informaci o typu závislosti mezi časovými řadami.

Bartlettův test

Bartlettův test se týká testu významnosti korelačního koeficientu r_0, takže je vhodný zejména v případech, kdy se závislost mezi časovými řadami projevuje bez zpoždění. Nyní nechť jsou $\{Y_t\}$ a $\{Z_t\}$ stacionární, lineární a centrované náhodné posloupnosti:

$$Y_t = b_0 \eta_t + b_1 \eta_{t-1} + \ldots,$$

$$Z_t = \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + \dots,$$

kde

$$\sum_{k=0}^{\infty} b_k^2 < \infty, \sum_{k=0}^{\infty} \beta_k^2 < \infty$$

a $\{\eta_t\}$ a $\{\varepsilon_t\}$ jsou centrované normální procesy a var $\eta_t = \sigma_1^2$, var $\varepsilon_t = \sigma_2^2$, $\mathrm{cov}(\eta_t, \varepsilon_t) = \rho\sigma_1\sigma_2$, $\mathrm{cov}(\eta_s, \eta_t) = \mathrm{cov}(\varepsilon_s, \varepsilon_t) = \mathrm{cov}(\eta_s, \varepsilon_t) = 0$ pro $s \neq t$. Pro kovarianční funkce $R_Y(t)$ a $R_Z(t)$ procesů $\{Y_t\}$ a $\{Z_t\}$ platí

$$R_Y(t) = \sigma_1^2 \sum_{k=0}^{\infty} b_{m+k} b_k \qquad \text{pro } m \geq 0 \tag{5.2}$$

a

$$R_Z(t) = \sigma_2^2 \sum_{k=0}^{\infty} \beta_{m+k} \beta_k \qquad \text{pro } m \geq 0. \tag{5.3}$$

Položme

$$r = \frac{\frac{1}{n} \sum_{k=1}^{n} Y_k Z_k}{\sqrt{R_Y(0) R_Z(0)}}. \tag{5.4}$$

Veličina r je zjednodušením korelačního koeficientu r_0 a mnohem lépe se s ní počítá. Jsou-li procesy $\{Y_t\}$ a $\{Z_t\}$ nezávislé, pak platí

$$Er = 0,$$

$$\text{var } r = \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} EY_j Z_j Y_k Z_k}{n^2 R_Y(0) R_Z(0)}. \tag{5.5}$$

Z předpokládané nezávislosti řad $\{Y_t\}$ a $\{Y_t\}$ máme

$$EY_j Z_j Y_k Z_k = EY_j Y_k EZ_j Z_k,$$

a (5.5) můžeme přepsat následujícím způsobem

$$\text{var } r = \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} R_Y(j-k) R_Z(j-k)}{n^2 R_Y(0) R_Z(0)}. \tag{5.6}$$

Pro obyčejný výběrový korelační koeficient r^* počítaný z nezávislých dvojic $(U_1, V_1), \dots, (U_M, V_M)$ se stejným regulárním dvojrozměrným normálním rozdělením, které má nulový korelační koeficient, platí

$$Er^* = 0,$$

$$\text{var } r^* = \frac{1}{M} + O(M^{-3/2}). \tag{5.7}$$

Viz např. Cramér [6].

Bartlett usoudil, že by r a r^* mohly mít velmi podobné rozdělení, protože vzorce pro tyto veličiny jsou téměř stejné. Aby se dosáhlo alespoň asymptoticky shody prvních dvou momentů, položme

$$M = \frac{1}{\text{var } r}. \tag{5.8}$$

Prakticky provádíme Bartlettův test takto:

1. Vypočteme výběrový korelační koeficient r_0.
2. Nalezneme vhodný model pro řady $\{Y_t\}$ a $\{Z_t\}$ a spočítáme příslušné kovarianční funkce $R_Y(t)$ a $R_Z(t)$ dané modelem.
3. Vypočteme var r podle vzorce (5.6).
4. Vypočteme $M = 1/\text{var } r$.
5. Významnost r_0 testujeme stejně, jako by se jednalo o obyčejný korelační koeficient počítaný z M nezávislých dvojic pozorování.

Podrobnější informace a přesnější zdůvodnění je možné nalézt ve skriptech [3].

Korelace residuí

Nechť $\{Y_t\}$ a $\{Z_t\}$ jsou stacionární autoregresní posloupnosti řádů p_1 a p_2:

$$Y_t = a_1 Y_{t-1} + \ldots + a_{p_1} Y_{t-p_1} + \eta_t, \tag{5.9}$$

kde η_t jsou nezávislé náhodné veličiny se stejným rozdělením $N(0, \sigma_1^2)$; $\sigma_1^2 > 0$ a

$$Z_t = \alpha_1 Z_{t-1} + \ldots + \alpha_{p_1} Z_{t-p_2} + \varepsilon_t, \tag{5.10}$$

kde ε_t jsou nezávislé náhodné veličiny se stejným rozdělením $N(0, \sigma_2^2)$; $\sigma_2^2 > 0$. Pokud bychom měli k dispozici realizaci procesů $\{\eta_t\}$ a $\{\varepsilon_t\}$, mohli bychom odhadnout jejich vzájemnou korelační funkci pomocí vzorce (5.1) jako

$$\tilde{r}_s = \frac{\sum\limits_{k=\max(1,1-s)}^{\min(n,n-s)} \eta_{s+k}\varepsilon_k}{\sqrt{\sum\limits_{k=1}^{n} \eta_k^2 \sum\limits_{k=1}^{n} \varepsilon_k^2}} \qquad \text{pro } s = 0, \pm 1, \pm 2, \ldots, \pm(n-1). \tag{5.11}$$

Jsou-li procesy $\{\eta_t\}$ a $\{\varepsilon_t\}$ nezávislé, pak při pevně zvoleném h má veličina

$$S_h^* = n \sum_{s=-h}^{h} \tilde{r}_s^2 \tag{5.12}$$

asymptoticky rozdělení χ^2_{2h+1}. Podrobněji viz skripta [3].

V praxi máme k dispozici pouze odhady residuí $\hat{\eta}_t$ a $\hat{\varepsilon}_t$, počítané rekurentně ze vztahů (5.9) a (5.10). Nabízí se vzít místo $\tilde{r}_s$ odhad

$$\hat{r}_s = \frac{\sum_{k=\max(1,1-s)}^{\min(n,n-s)} \hat{\eta}_{s+k}\hat{\varepsilon}_k}{\sqrt{\sum_{k=1}^{n} \hat{\eta}_k^2 \sum_{k=1}^{n} \hat{\varepsilon}_k^2}} \qquad \text{pro } s = 0, \pm 1, \pm 2, \ldots, \pm(n-1). \tag{5.13}$$

Haugh [11] dokázal, že i pak má veličina

$$S_h = n \sum_{s=-h}^{h} \hat{r}_s^2 \tag{5.14}$$

při $n \to \infty$ asymptoticky rozdělení χ^2_{2h+1}.

5.2 Spektrální metody

Závislost mezi časovými řadami můžeme vyšetřovat také pomocí spektrální analýzy vícerozměrných časových řad.

Jsou-li $\{X_t^1; t \in T\}, \ldots, \{X_t^p; t \in T\}$ reálné náhodné posloupnosti (tj. T je množina celých čísel) s konečnými druhými momenty a středními hodnotami $\mu_t^1, \ldots, \mu_t^p$, zavádíme vzájemnou kovarianční funkci j-té a k-té náhodné posloupnosti vztahem

$$R_{jk}(s,t) = E(X_s^j - \mu_s^j)(X_t^k - \mu_t^k) \qquad \text{pro } j,k = 1,2,\ldots,p. \tag{5.15}$$

Jsou-li střední hodnoty $\mu_t^1, \ldots, \mu_t^p$ konstantní a závisí-li všechny vzájemné kovarianční funkce $R_{jk}(s,t)$ na svých argumentech pouze prostřednictvím jejich rozdílu, říkáme že p-rozměrný náhodný proces $\{(X_t^1, \ldots, X_t^p); t \in T\}$ je (slabě) stacionární. Podobně jako v jednorozměrném případě se pak místo $R_{jk}(s,t)$ píše $R_{jk}(s-t)$.

Pokud platí, že $R_{jj}(0) > 0$ pro všechna $j = 1, \ldots, p$, zavádí se vzájemné korelační funkce

$$B_{jk}(t) = \frac{R_{jk}(t)}{\sqrt{R_{jj}(0)R_{kk}(0)}}. \tag{5.16}$$

Spektrální rozklad vzájemné kovarianční funkce

Dá se dokázat (viz např. [1]), že každou vzájemnou kovarianční funkci $R_{jk}(t)$ lze vyjádřit ve tvaru

$$R_{jk}(t) = \int_{-\pi}^{\pi} e^{it\lambda}\, dF_{jk}(\lambda), \tag{5.17}$$

kde $F_{jk}(\lambda)$ jsou spektrální distribuční funkce. Všechny funkce $F_{jk}(\lambda)$ jsou zleva spojité. Funkce $F_{jj}(\lambda)$ jsou navíc reálné a $F_{jj}(-\pi) = 0$, $F_{jj}(\pi) = R_{jj}(0)$.

Jsou-li všechny funkce $F_{jk}(\lambda)$ absolutně spojité a označíme-li $F'_{jk}(\lambda) = f_{jk}(\lambda)$, můžeme přepsat (5.17) ve tvaru

$$R_{jk}(t) = \int_{-\pi}^{\pi} e^{it\lambda} f_{jk}(\lambda)\ d\lambda. \tag{5.18}$$

Funkcím $f_{jk}(\lambda)$ se říká spektrální hustoty. Dá se dokázat, že spektrální hustoty existují, pokud platí podmínka

$$\sum_{t=-\infty}^{\infty} |R_{jj}(t)| < \infty \qquad \text{pro } j = 1, 2, \ldots, p. \tag{5.19}$$

Spektrální hustota $f_{jk}(\lambda)$ s různými indexy $j \neq k$ je obecně komplexní funkce a platí pro ni $f_{jk}(\lambda) = \overline{f_{kj}(\lambda)}$ a $f_{jk}(\lambda) = \overline{f_{jk}(-\lambda)}$.

Koherenční koeficient a fáze

Nyní předpokládejme, že všechny spektrální hustoty $f_{jk}(\lambda)$ pro $j, k = 1, \ldots, p$ existují a $f_{jj}(\lambda) > 0$ pro $j = 1, \ldots, p$ a $-\pi \leq \lambda \leq \pi$. Jako míra závislosti mezi odpovídajícími periodickými složkami procesů $\{X_t^j\}$ a $\{X_t^k\}$ se zavádí koherenční koeficient

$$C_{jk}(\lambda) = \frac{|f_{jk}(\lambda)|}{\sqrt{f_{jj}(\lambda) f_{kk}(\lambda)}} \qquad \text{pro } \lambda \in < -\pi, \pi > . \tag{5.20}$$

Tento vzorec připomíná definici korelačního koeficientu a také interpretace korelačního a koherenčního koeficientu je podobná. Hodnocení koherenčního koeficientu závisí na metodě jeho odhadu.

Pro nezávislé složky je příslušný koherenční koeficient nulový. Dále se dá dokázat (opět viz [1]), že

$$0 \leq C_{jk}(\lambda) \leq 1 \qquad \text{pro } \lambda \in < -\pi, \pi > . \tag{5.21}$$

Nechť $c_{jk}(\lambda)$ je reálná část spektrální hustoty $f_{jk}(\lambda)$ a $q_{jk}(\lambda)$ je její imaginární část. Potom

$$f_{jk}(\lambda) = c_{jk}(\lambda) + i q_{jk}(\lambda). \tag{5.22}$$

Funkci $c_{jk}(\lambda)$ se říká kospektrum (anglicky „cospectral density“) a funkci $q_{jk}(\lambda)$ kvadratické spektrum („quadrature spectral density“). Pomocí $c_{jk}(\lambda)$ a $q_{jk}(\lambda)$ můžeme koherenční koeficient napsat ve tvaru

$$C_{jk}(\lambda) = \frac{\sqrt{c_{jk}^2(\lambda) + q_{jk}^2(\lambda)}}{\sqrt{f_{jj}(\lambda) f_{kk}(\lambda)}} \qquad \text{pro } \lambda \in < -\pi, \pi > . \tag{5.23}$$

Je-li kospektrum $c_{jk}(\lambda) \neq 0$, definuje se fáze

$$\Phi_{jk}(\lambda) = \arctan\left[\frac{q_{jk}(\lambda)}{c_{jk}(\lambda)}\right]. \tag{5.24}$$

Ve fázovém koeficientu se promítá časové zpoždění odpovídajících periodických složek. Předpokládejme, že každá periodická složka s frekvencí λ řady X_t^k je rovna periodické složce se stejnou frekvencí λ řady X_t^j zpožděné o hodnotu $\varphi(\lambda)$. Pak lze ukázat (viz [5] a [1]), že $C_{jk}(\lambda) = 1$ a $\Phi_{jk}(\lambda) = \arctan[\tan\varphi(\lambda)]$. Je-li $-\pi/2 < \varphi(\lambda) < \pi/2$, pak se fázový koeficient $\Phi_{jk}(\lambda)$ rovná přímo fázovému posunu $\varphi(\lambda)$. Je-li např. $\pi/2 < \varphi(\lambda) < 3\pi/2$, pak $\Phi_{jk}(\lambda) = \varphi(\lambda) - \pi$ atd. Časové zpoždění řady X_t^k za řadou X_t^j v rámci frekvence λ dostaneme jako $\varphi(\lambda)/\lambda$.

Odhad vzájemného spektra, koherenčního koeficientu a fáze

Pro zjednodušení předpokládejme, že máme k dispozici konečnou realizaci dvou časových řad $Y_1, \ldots, Y_n$ a $Z_1, \ldots, Z_n$. Kovarianční funkci $\{Y_t\}$ budu značit $R_Y(t)$, kovarianční funkci $\{Z_t\}$ $R_Z(t)$ a vzájemnou kovarianční funkci procesů $\{Y_t\}$ a $\{Z_t\}$ budu značit $R_{YZ}(t)$. Zcela obdobně budu indexovat i příslušné spektrální hustoty, koherenční koeficient, fázi a jejich odhady.

Vzájemná spektrální hustota $f_{YZ}(\lambda)$ se odhaduje zcela obdobně jako obyčejná spektrální hustota.

Nechť náhodná posloupnost $\{(Y_t, Z_t)\}$ je normální a centrovaná. Vzájemný periodogram $\{Y_t\}$ a $\{Z_t\}$ je definován jako

$$I_{jk}(\lambda) = \frac{1}{2\pi n} \sum_{t=1}^{n} Y_t e^{-it\lambda} \sum_{s=1}^{n} Z_s e^{is\lambda} \qquad \text{pro } \lambda \in < -\pi, \pi > . \tag{5.25}$$

K „rozumnému“ odhadu vzájemné spektrální hustoty $f_{YZ}(\lambda)$ se dojde obdobně jako při odhadu obyčejné spektrální hustoty. Jako odhad $f_{YZ}(\lambda)$ dostaneme

$$\hat{f}_{YZ}(\lambda) = \sum_{t=-n+1}^{n-1} w_t c_{YZ}(\lambda) e^{-it\lambda}, \tag{5.26}$$

kde w_t je váhová funkce (nejčastěji se používají Parzenovy váhy (4.26)) a $c_{YZ}(t)$ je odhad vzájemné kovarianční funkce

$$\begin{array}{lcll} c_{YZ}(t) & = & \frac{1}{n} \sum_{s=1}^{n-t} Y_{s+t} Z_s & \text{pro } t = 0, 1, \ldots, n-1, \\ c_{YZ}(t) & = & c_{ZY}(-t) & \text{pro } t = -n+1, \ldots, -2, -1. \end{array} \tag{5.27}$$

Rozepsáním vzorce (5.26) na reálnou a imaginární část dostaneme následující odhady pro kospektrum

$$\hat{c}_{YZ}(\lambda) = w_0 + \sum_{t=1}^{n-1} w_t [c_{YZ}(t) + c_{ZY}(t)] \cos t\lambda \tag{5.28}$$

a kvadratické spektrum

$$\hat{q}_{YZ}(\lambda) = \sum_{t=1}^{n-1} w_t [c_{ZY}(t) - c_{YZ}(t)] \sin t\lambda. \tag{5.29}$$

Za odhad koherenčního koeficientu $C_{YZ}(\lambda)$ se bere výběrový koherenční koeficient

$$\hat{C}_{YZ}(\lambda) = \frac{\sqrt{\hat{c}_{YZ}^2(\lambda) + \hat{q}_{YZ}^2(\lambda)}}{\sqrt{\hat{f}_Y(\lambda)\hat{f}_Z(\lambda)}}. \tag{5.30}$$

Pokud při odhadu spektrálních hustot použijeme Parzenovy váhy, platí

$$0 \leq \hat{C}_{YZ}(\lambda) \leq 1, \tag{5.31}$$

při použití jiných vah není tato vlastnost odhadu koherenčního koeficientu zaručena. Rozdělením výběrového koherenčního koeficientu se zabýval Goodman [8] [9]. Ukázal, že rozdělení $\hat{C}_{YZ}(\lambda_0)$ lze za předpokladu $C_{YZ}(\lambda_0) = 0$ aproximovat rozdělením s distribuční funkcí

$$F(u) = 1 - (1 - u^2)^{b-1}, \tag{5.32}$$

kde b je tzv. ekvivalentní počet stupňů volnosti. Při použití Parzenových vah je $b = 1.8n/m$.

Fáze $\Phi_{YZ}(\lambda)$ se odhaduje pomocí

$$\hat{\Phi}_{YZ}(\lambda) = \arctan\left[\frac{\hat{q}_{YZ}(\lambda)}{\hat{c}_{YZ}(\lambda)}\right]. \tag{5.33}$$

Kapitola 6

Wolfova čísla

Wolfova čísla jako míru sluneční aktivity zavedl profesor Rudolf Wolf z Curychu v roce 1849. Definoval je jako $W = K(10g + f)$, kde g je počet skupin slunečních skvrn, f je celkový počet slunečních skvrn a K je konstanta pro danou observatoř. Profesor Wolf je zpětně z nejrůznějších pramenů určil pro roky 1749 až 1848.

Časová řada ročních průměrů Wolfových čísel je asi nejčastěji analyzovanou časovou řadou vůbec. Některé přístupy k předpovídání slunečního cyklu využivají vztahy mezi velikostí maxima, dobou růstu, dobou poklesu a ostatními charakteristickými rysy slunečního cyklu. Např. Waldmeier (1968) roztřídil cykly podle jejich vlastností a navrhl tabulku pro předpověď sluneční aktivity založenou na „standardních“ cyklech. Yule [29] použil pro Wolfova čísla autoregresi druhého řádu, ale Moran [17] ukázal, že model AR(2) pro časovou řadu ročních průměrů Wolfových čísel od roku 1751 do roku 1950, který odhadl jako

$$x_t = 14.524 + 1.350x_{t-1} - 0.6613x_{t-2}$$

neposkytuje uspokojivé předpovědi. Schaerf [21] použil signifikantní koeficienty a pro vycentrovanou časovou řadu ročních průměrů Wolfových čísel od roku 1749 do roku 1924 dostal model

$$z_t = 1.239z_{t-1} - 0.55z_{t-2} + 0.128z_{t-9}.$$

Box a Jenkins [4] uvažovali obecnější ARMA modely a došli také k autoregresi druhého řádu

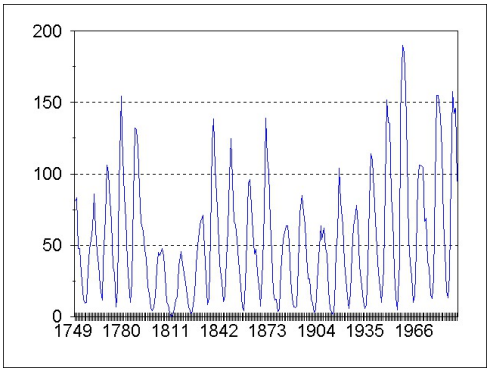
$$x_t = 14.9 + 1.32x_{t-1} - 0.63x_{t-1}$$

pro časovou řadu ročních průměrů Wolfových čísel od roku 1770 do roku 1869. Morris [18] použil krokovou regresi na autoregresní model

$$x_t = a_0 + a_1x_{t-1} + \ldots + a_px_{t-p} + \varepsilon_t$$

pro $p = 30$, a pro roční průměry Wolfových čísel od roku 1755 do roku 1964 dostal autoregresní model

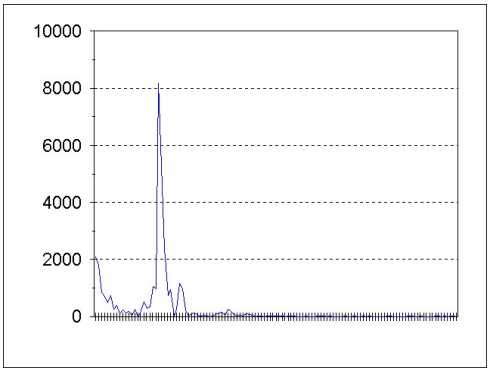
$$x_t = 5.055 + 1.250x_{t-1} - 0.538x_{t-2} + 0.189x_{t-9}.$$



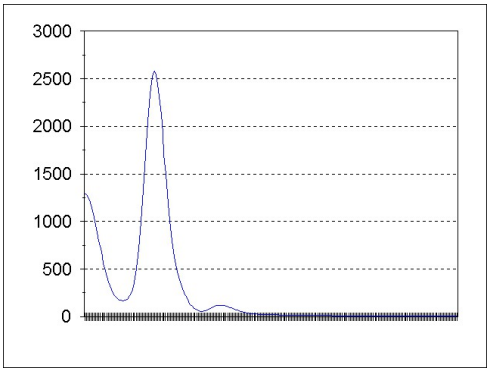
Obrázek 6.1: Roční průměry Wolfových čísel

Roční průměry Wolfových čísel z let 1749 až 1992 jsou uvedeny v následující tabulce. Stejná data jsou graficky znázorněna na obrázku 6.1. Dále budu pracovat pouze s vycentrovanými daty. Výběrový průměr Wolfových čísel je 52.45656. Centrovanou časovou řadu ročních průměrů Wolfových čísel od roku 1749 do roku 1992 budu v dalším textu značit $W_1, \ldots, W_{244}$.

1749	80.9	83.4	47.7	47.8	30.7	12.2	9.6	10.2	32.4	47.6
1759	54.0	62.9	85.8	61.2	45.1	36.4	20.9	11.4	37.8	69.8
1769	106.1	100.8	81.6	66.5	34.8	30.6	7.0	19.8	92.5	154.4
1779	126.2	84.8	68.1	38.5	22.8	10.2	24.1	82.9	132.0	130.9
1789	118.1	89.9	66.6	60.0	46.9	41.0	21.3	16.0	6.4	4.1
1799	6.8	14.5	34.0	45.0	43.0	47.5	42.2	28.1	10.1	8.1
1809	2.5	0.0	1.4	5.0	12.2	13.9	35.4	45.8	41.1	30.1
1819	24.0	15.6	6.6	4.0	1.8	8.6	16.6	36.6	49.6	64.2
1829	67.0	70.9	47.8	27.5	8.5	13.2	56.9	121.5	138.3	103.2
1839	85.7	64.7	36.7	24.2	10.7	15.0	40.1	61.5	98.4	124.7
1849	96.3	66.6	64.5	54.1	39.0	20.6	6.7	4.3	22.7	54.8
1859	93.8	95.8	77.2	59.1	44.0	47.0	30.5	16.3	7.3	37.6
1869	74.0	139.0	111.2	101.6	66.2	44.7	17.0	11.3	12.4	3.4
1879	6.0	32.2	54.3	59.6	63.6	63.5	52.0	25.4	13.1	6.8
1889	6.2	7.1	35.6	72.9	85.1	78.0	64.0	41.8	26.2	26.7
1899	12.1	9.5	2.7	5.4	24.4	42.0	63.5	53.9	62.0	48.5
1909	43.9	18.6	5.7	3.6	1.4	9.6	47.4	57.1	103.9	80.6
1919	63.6	37.6	26.1	14.2	5.8	16.7	44.3	63.9	69.0	77.8
1929	64.9	35.7	21.2	11.1	5.7	8.7	36.0	79.7	114.4	109.6
1939	88.8	67.8	47.5	30.6	16.1	9.8	33.1	92.5	151.5	136.2
1949	135.1	83.9	69.4	31.4	13.9	4.4	38.0	141.7	189.9	184.6
1959	158.8	112.3	53.9	37.6	27.9	10.2	15.1	46.9	93.7	105.9
1969	105.6	104.7	66.7	68.9	38.2	34.4	15.5	12.6	27.5	92.7
1979	155.3	154.7	140.4	116.3	66.6	45.5	17.9	13.4	29.2	100.0
1989	157.8	142.3	145.8	94.5						



Obrázek 6.2: Periodogram pro Wolfova čísla



Obrázek 6.3: Odhad spektrální hustoty pro Wolfova čísla

Testy náhodnosti

Přestože už na první pohled Wolfova čísla rozhodně nevypadají jako posloupnost nezávislých náhodných veličin, provedl jsem všechny testy náhodnosti popsané v teoretické části.

V časové řadě ročních průměrů Wolfových čísel je 99 bodů růstu. Střední hodnota počtu bodů růstu je 121.5, rozptyl je 20.42. Testová statistika $\frac{|99-121.5|}{\sqrt{20.42}} = |-4.9796| > 1.96$ a na pětiprocentní hladině pravděpodobnosti hypotézu náhodnosti dat zamítáme kvůli příliš malému počtu bodů růstu.

Počet bodů zvratu je 61. Testová statistika $\frac{|61-161.33|}{\sqrt{43.06}} = |-15.2908| > 1.96$ a náhodnost dat tedy zamítá i test založený na bodech zvratu.

Spearmanův koeficient pořadové korelace je 0.1511 a opět je překročena kritická hodnota na pětiprocentní hladině pravděpodobnosti.

Při mediánovém testu leží nad mediánem 122 pozorování, pod mediánem 121 pozorování a na mediánu jedno pozorování (medián je 43.9). Počet vytvořených sérií je 47. Testová statistika nabývá hodnoty -9.5781 a je překročena kritická hodnota. Počet vytvořených skupin je výrazně menší než by měl být.

Test pomocí odhadu rozptylu ze sousedních pozorování náhodnost dat na pětiprocentní

hladině pravděpodobnosti také zamítá (hodnota statistiky d je 0.3588), protože sousední pozorování jsou bližší než plyne z předpokladu náhodnosti.

Fisherův test je založen na periodogramu, který je na obrázku 6.2. Nejvyšší hodnota periodogramu je dosažena pro frekvenci odpovídající délce periody 11.09. Hodnota statistiky W z Fisherova testu je $W = 0.2431$ a je překročena příslušná kritická hodnota (0.0642). Pomocí postupu, který navrhl Whittle pro testování významnosti dalších periodických složek dostaneme postupně periody: 10.61, 10.17, 244, 122, 8.41, 12.2, 8.13, 11.62, 9.38, 81.33, 9.76, 40.67, 61, 14.35, 48.8. Hodnota statistiky T ze Siegelova testu je $T = 0.3730$ a kritická hodnota (0.0327) je také překročena. Odhad spektrální hustoty je na obrázku 6.3. V následující tabulce jsou podrobně uvedeny výsledky Fisherova testu.

délka periody	statistika W	aproximace kritické hodnoty
11.09	0.2431	0.0642
10.61	0.1790	0.0646
10.17	0.1144	0.0651
244	0.1128	0.0656
122	0.1101	0.0661
8.41	0.0799	0.0666
12.20	0.0789	0.0671
8.13	0.0797	0.0676
11.62	0.0854	0.0681
9.38	0.0915	0.0686
81.33	0.0902	0.0692
9.76	0.0874	0.0697
40.67	0.0935	0.0703
61.0	0.1011	0.0709
14.35	0.0817	0.0714
48.80	0.0861	0.0720

Výsledky testů náhodnosti jenom potvrzují, co je na obrázku 6.1 vidět na první pohled.

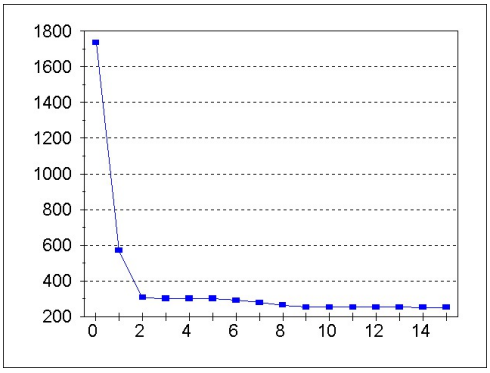
Autoregrese

Prvním problémem při konstrukci autoregresního je hledání správného řádu modelu. K tomuto účelu můžeme využít odhady autokorelací a parciálních autokorelací, které jsou uvedeny v následující tabulce.

	1	2	3	4	5	6	7	8
autokorelace	0.817	0.443	0.033	-0.275	-0.411	-0.352	-0.136	0.169
parciální autokorelace	0.817*	-0.679*	-0.111	0.036	0.024	0.146*	0.184*	0.229*

	9	10	11	12	13	14	15	16
autokorelace	0.462	0.628	0.608	0.420	0.146	-0.108	-0.272	-0.317
parciální autokorelace	0.191*	0.016	-0.004	0.028	0.011	0.064	-0.050	-0.104

Významné hodnoty parciální autokorelační funkce jsou označeny hvězdičkou. Identifikační bod parciální autokorelační funkce je zřejmě 9.



Obrázek 6.4: Odhad rozptylu bílého šumu pro Wolfova čísla jako AR(k)

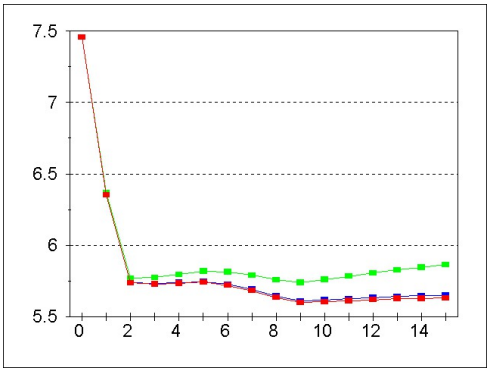
Na obrázku 6.4 jsou odhady rozptylu bílého šumu po autoregresi k-tého řádu. Odhad rozptylu bílého šumu výrazně klesne pro model AR(2) a dále se už příliš nemění. Po optickém posouzení obrázku 7.2 bych zřejmě zvolil model AR(2).

Nejobjektivnější nástroj pro volbu řádu autoregrese jsou informační kritéria popsaná v teoretické části. Hodnoty kritérií AIC, BIC a HQ jsou uvedeny v následující tabulce. Stejné hodnoty jsou graficky znázorněny na obrázku 6.5. Všechna kritéria určují jako nejlepší model autoregresi řádu 9. Nejmenší hodnoty kritérií jsou v tabulce označeny hvězdičkou.

řád AR modelu	odhad σ_ε^2	AIC	BIC	HQ
0	1733.3956	7.4578	7.4578	7.4578
1	569.9029	6.3537	6.3680	6.3525
2	305.9278	5.7397	5.7684	5.7373
3	300.9487	5.7315	5.7745	5.7279
4	300.5520	5.7384	5.7957	5.7336
5	300.5304	5.7465	5.8182	5.7405
6	292.1088	5.7263	5.8123	5.7190
7	279.6277	5.6908	5.7912	5.6824
8	264.6641	5.6440	5.7587	5.6343
9	253.4878	5.6091*	5.7381*	5.5982*
10	253.4579	5.6172	5.7605	5.6050
11	253.4506	5.6253	5.7830	5.6120
12	253.4196	5.6334	5.8054	5.6189
13	253.4182	5.6416	5.8279	5.6258
14	252.1320	5.6447	5.8454	5.6277
15	251.3096	5.6496	5.8646	5.6315

Vzhledem ke shodě všech informačních kritérií jsem se rozhodl pro model AR(9). Odhad parametrů autoregresního modelu řádu 9 pro Wolfova čísla podmíněnou metodou nejmenších čtverců je

$$X_t = 1.1642X_{t-1} - 0.4056X_{t-2} - 0.1542X_{t-3} + 0.1280X_{t-4} - 0.0933X_{t-5} +$$



Obrázek 6.5: AIC, BIC a HQ pro stanovení řádu autoregrese Wolfových čísel

$$+0.0335X_{t-6} + 0.0048X_{t-7} - 0.0209X_{t-8} + 0.2157X_{t-9} + \varepsilon_t, \qquad (6.1)$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma^2 = 253.488$. Hodnota portmanteau statistiky je $Q_{17} = 7.9470$, hodnota Box-Pierceovy statistiky je $Q^*_{17} = 8.4738$, příslušná kritická hodnota rozdělení χ^2 o 8 stupních volnosti je 27.9, hodnota McLeod-Liovy statistiky je 18.7796 a příslušná kritická hodnota je 42.9. Adekvátnost tohoto modelu tedy nezamítá žádný test popsaný v teoretické části. Konfidenční intervaly pro autoregresní parametry, hodnota t-statistiky pro test hypotézy, že daný parametr je nula, a nejmenší hladina pravděpodobnosti, na které zamítáme nulovost parametru jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	1.1642	(1.0384,1.2900)	18.1396	0.0000
2	-0.4056	(-0.6030,-0.2081)	-4.0256	0.0001
3	-0.1542	(-0.3622,0.0536)	-1.4540	0.1459
4	0.1280	(-0.0837,0.3397)	1.1853	0.2359
5	-0.0933	(-0.3050,0.1184)	-0.8636	0.3878
6	0.0335	(-0.1793,0.2463)	0.3084	0.7578
7	0.0048	(-0.2078,0.2173)	0.0440	0.8387
8	-0.0209	(-0.2257,0.1839)	-0.2001	0.8414
9	0.2157	(0.0844,0.3470)	3.2189	0.0013

Odhad parametrů podmíněnou metodou nejmenších čtverců úzce souvisí s detekcí odlehlých pozorování pomocí diagonálních prvků projekční matice H. Touto metodou dostaneme odlehlé vektory pro $t = 31, 32, 33, 34, 35, 36, 37, 38, 124, 125, 126, 127, 128, 129, 130, 203, 204, 205, 206, 207, 210, 211, 212, 213, 214, 215, 216, 217, 218, 232, 237, 243, 244$. Jako odlehlá pozorování můžeme tedy identifikovat pozorování W_{30}, W_{123}, W_{202} a W_{209}. Tato pozorování odpovídají letům 1778, 1871, 1950 a 1957.

U šesti (z celkového počtu devíti) parametrů modelu AR(9) nezamítáme na pětiprocentní hladině pravděpodobnosti jejich nulovost. Proto jsem se pokusil najít vhodnější model, ve kterém W_t závisí pouze na nějaké podmnožině z veličin $W_{t-1}, \ldots, W_{t-K}$, kde K je maximální počet parametrů. Vhodnost těchto modelů můžeme (stejně jako vhodnost obyčejných autoregresních modelů) posoudit pomocí informačních kritérií AIC, BIC a HQ. Zvolil jsem $K = 15$ a hledal model, pro který jsou AIC, BIC a HQ nejmenší. Všechna kritéria shodně

určila model

$$W_t = 1.2048W_{t-1} - 0.5163W_{t-2} + 0.2072W_{t-9} + \varepsilon_t, \tag{6.2}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 256.6822$. Odhad rozptylu bílého šumu se oproti modelu AR(9) zvětšil pouze nepatrně a přitom se třikrát snížil počet parametrů. Hodnota portmanteau statistiky je $Q_{17} = 9.9360$, hodnota Box-Pierceovy statistiky je $Q_{17}^* = 10.4322$, příslušná kritická hodnota rozdělení χ^2 o 14 stupních volnosti je 38.1, hodnota McLeod-Liovy statistiky je $Q_{17}^{**} = 18.7796$ a příslušná kritická hodnota je 42.9. Ani jeden test tedy nezamítá adekvátnost tohoto modelu. Konfidenční intervaly pro autoregresní parametry jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	1.2048	(1.1056,1.3040)	23.8009	0.0000
2	-0.5163	(-0.6159,0.4167)	-10.1583	0.0000
9	0.2072	(0.1475,0.2669)	6.7998	0.0000

Kapitola 7

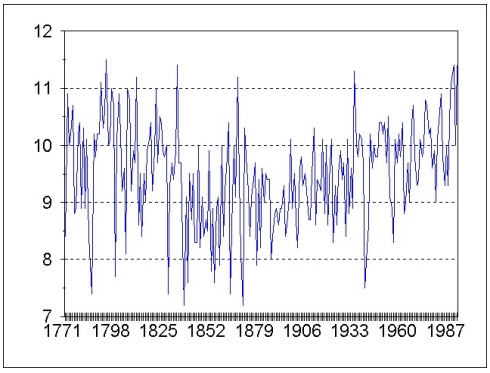
Teploty

V následující tabulce jsou uvedeny roční průměrné teploty v Klementinu od roku 1771 do roku 1992. Stejné hodnoty jsou graficky znázorněny na obrázku 7.1.

1771	8.4	10.9	10.0	10.2	10.7	8.8	8.9	10.2	10.4	8.9
1781	10.3	8.9	10.1	8.4	7.9	7.4	10.2	9.8	10.2	10.2
1791	11.1	10.3	10.6	11.5	10.0	10.1	11.0	10.7	7.7	10.1
1801	10.9	10.2	9.2	9.6	8.1	11.0	10.8	9.2	9.9	9.7
1811	11.2	8.6	9.5	8.4	9.5	9.0	9.9	10.1	10.4	9.2
1821	9.9	11.0	9.7	10.5	10.4	9.9	9.8	10.0	7.4	9.3
1831	9.7	9.4	9.9	11.4	9.7	9.7	8.3	7.2	9.1	7.6
1841	9.5	8.7	9.5	8.3	8.3	10.0	8.2	9.1	8.4	8.7
1851	8.5	9.9	7.8	8.9	7.6	8.9	9.1	7.9	10.0	8.4
1861	9.2	9.8	10.4	7.4	9.1	10.0	9.1	11.2	9.4	8.1
1871	7.2	10.3	9.8	9.2	8.4	9.1	9.4	9.7	7.9	9.4
1881	8.2	9.6	9.0	9.5	9.4	9.4	8.0	8.4	8.8	8.9
1891	8.6	8.9	8.9	9.3	8.4	8.6	9.1	10.1	8.9	9.5
1901	8.7	8.2	9.6	9.8	9.3	9.5	9.2	8.7	8.7	9.5
1911	10.3	8.6	9.4	9.3	9.2	10.1	8.8	10.0	8.6	9.7
1921	10.1	8.3	9.3	8.6	9.5	9.9	9.4	9.7	8.4	10.1
1931	8.8	9.6	8.9	11.3	10.0	9.8	10.2	10.1	9.7	7.5
1941	8.2	8.6	10.2	9.6	10.0	9.8	9.8	10.4	10.4	10.2
1951	10.4	9.7	10.5	9.1	8.9	8.3	10.1	9.7	10.2	9.8
1961	10.4	8.8	9.0	9.7	9.0	10.4	10.7	9.8	9.3	9.4
1971	10.1	9.8	10.0	10.8	10.7	10.2	10.3	9.6	9.9	9.0
1981	10.1	10.6	10.9	9.8	9.3	10.0	9.3	10.9	11.2	11.4
1991	10.0	11.4								

Nejdříve jsem na data použil všechny testy náhodnosti popsané v teoretické části.

V časové řadě ročních průměrných teplot je 113 bodů růstu. Po vyškrtnutí stejných sousedních hodnot v řadě zůstalo 214 pozorování. Střední hodnota počtu bodů růstu je 106.5, rozptyl je 17.92. Testová statistika $\frac{|113-106.5|}{\sqrt{17.92}} = 1.54 < 1.96$ a na pětiprocentní hladině pravděpodobnosti hypotézu náhodnosti dat nezamítáme.



Obrázek 7.1: Průměrné roční teploty v Klementinu

Počet bodů zvratu je 150. Testová statistika $\frac{|150-141.33|}{\sqrt{37.72}} = 1.41 < 1.96$ a náhodnost dat tedy nezamítá ani test založený na bodech zvratu.

Spearmanův koeficient pořadové korelace je 0.0951 a opět není překročena kritická hodnota na pětiprocentní hladině pravděpodobnosti.

Při mediánovém testu leží nad mediánem 107 pozorování, pod mediánem 109 pozorování a na mediánu 6 pozorování (medián je 9.6). Počet vytvořených sérií je 76. Testová statistika nabývá hodnoty -4.36 a je překročena kritická hodnota. Počet vytvořených skupin je menší než by měl být. To by mohlo být způsobeno například přítomností rostoucí nebo klesající tendence nebo přítomností periodické složky s dlouhou periodou.

Test pomocí odhadu rozptylu ze sousedních pozorování náhodnost dat na pětiprocentní hladině pravděpodobnosti také zamítá (hodnota statistiky d je 1.48), protože sousední pozorování jsou bližší než plyne z předpokladu náhodnosti.

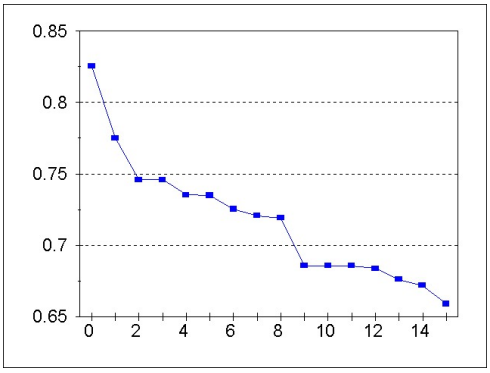
Výsledky testů náhodnosti ukazují na přítomnost rostoucí nebo klesající tendence nebo periodické složky. Vzhledem k tomu, že náhodnost dat nezamítly testy založené na bodech růstu a na Spearmanově koeficientu, bych se přikláněl spíše k té periodické složce. Tuto domněnku potvrzují i výsledky Fisherova a Siegelova testu, které jsou uvedeny dále.

Dále budu pracovat pouze s vycentrovanými daty. Výběrový průměr této časové řady je 9.4937. Centrovanou časovou řadu průměrných ročních teplot od roku 1771 do roku 1992 budu v dalším textu značit jako $T_1, \ldots, T_{222}$.

7.1 Autoregrese

Nejdříve je zapotřebí určit vhodný řád autoregrese. Odhady autokorelační a parciální autokorelační funkce r_k a r_{kk} pro $k = 1, \ldots, 16$ jsou uvedeny v následující tabulce.

	1	2	3	4	5	6	7	8
autokorelace	0.244	0.240	0.102	0.167	0.089	0.171	0.147	0.067
parciální autokorelace	0.244*	0.192*	0.009	0.109	0.017	0.108	0.077	-0.044
	9	10	11	12	13	14	15	16
autokorelace	0.242	0.118	0.122	0.016	0.166	0.124	0.238	0.189
parciální autokorelace	0.210*	-0.000	-0.001	-0.054	0.103	0.077	0.126	0.068



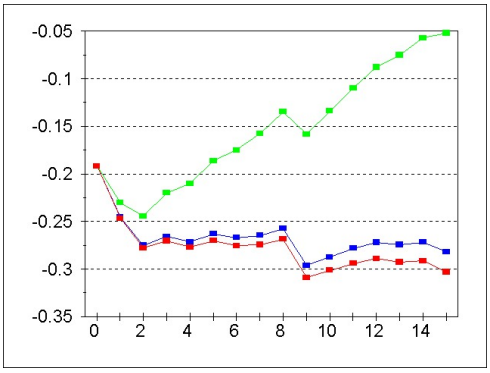
Obrázek 7.2: Odhad rozptylu bílého šumu pro teploty jako AR(k)

K určení řádu autoregrese slouží především odhady parciální autokorelace r_{kk}, které by se pro k větší než je správný řád autoregrese neměly významně lišit od nuly. Významnost r_{kk} můžeme posoudit pomoci Quenouilleovy aproximace (1.9) pro směrodatnou odchylku r_{kk}. V našem případě je $n = 222$ a tedy $\hat{\sigma}(r_{kk}) = 1/\sqrt{222} = 0.067$. Jednoduchý test významnosti odhadů parciální autokorelační funkce získáme jejich porovnáním s hodnotou $2 \cdot 0.067 = 0.134$. Významné hodnoty jsou v tabulce označeny hvězdičkou. Výsledek tohoto postupu ukazuje, že by se měla použít autoregrese řádu 2 nebo 9.

K určení řádu autoregrese mohou posloužit i odhady rozptylu bílého šumu na obrázku 7.2. Vidíme, že odhady kolísají přibližně kolem stejné hodnoty pro řády 2 až 8 a u devítky ještě o něco klesnou. Tato metoda tedy také ukazuje na model AR(2) nebo AR(9).

Další a nejobjektivnější možnost je použití informačních kritérií. V následující tabulce jsou uvedeny hodnoty odhadu rozptylu bílého šumu, AIC, BIC a HQ.

řád AR modelu	odhad σ_ε^2	AIC	BIC	HQ
0	0.8255	-0.1918	-0.1918	-0.1918
1	0.7752	-0.2456	-0.2303	-0.2470
2	0.7460	-0.2750	-0.2443*	-0.2778
3	0.7459	-0.2661	-0.2202	-0.2704
4	0.7353	-0.2714	-0.2101	-0.2771
5	0.7348	-0.2631	-0.1864	-0.2701
6	0.7254	-0.2670	-0.1750	-0.2755
7	0.7206	-0.2646	-0.1574	-0.2745
8	0.7192	-0.2575	-0.1348	-0.2687
9	0.6857	-0.2963*	-0.1583	-0.3090*
10	0.6856	-0.2873	-0.1340	-0.3014
11	0.6856	-0.2784	-0.1098	-0.2939
12	0.6838	-0.2720	-0.08819	-0.2889
13	0.6760	-0.2745	-0.0752	-0.2928
14	0.6717	-0.2717	-0.0572	-0.2915
15	0.6589	-0.2820	-0.0521	-0.3031



Obrázek 7.3: AIC, BIC a HQ pro stanovení řádu autoregrese teplot

Minimální hodnoty jednotlivých kritérií jsou v tabulce označeny hvězdičkou. Kritéria AIC a HQ se shodují na modelu AR(9), kritérium BIC určilo model AR(2). Hodnoty informačních kritérií jsou graficky znázorněny na obrázku 7.3.

Nakonec jsem se rozhodl pro průměrné roční teploty v Klementinu použít model AR(9). Pomocí podmíněné metody nejmenších čtverců popsané v teoretické části jsem dostal následující model:

$$T_t = 0.1957T_{t-1} + 0.1432T_{t-2} - 0.0395T_{t-3} + 0.1044T_{t-4} - 0.0282T_{t-5} + \\ +0.1128T_{t-6} + 0.0590T_{t-7} - 0.0807T_{t-8} + 0.2216T_{t-9} + \varepsilon_t, \qquad (7.1)$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma^2 = 0.6857$.

Model (7.1) jsem ověřil všemi metodami popsanými v teoretické části. Hodnota portmanteau statistiky Q pro $m = 16$ je $Q_{16} = 8.3396$, příslušná kritická hodnota na pětiprocentní hladině pravděpodobnosti je 26.0. Hodnota Box-Pierceovy statistiky Q^*_{16} je 8.9622, kritická hodnota je stejná jako pro portmanteau statistiku. Portmanteau ani Box-Pierceova statistika platnost ověřovaného modelu nezamítají. McLeod-Liova statistika $Q^{**}_{16} = 17.3851$ a příslušná kritická hodnota je 41.3, takže ani tento test nezamítá platnost modelu AR(9).

Z programu SOLO jsem dostal následující tabulku, ve které jsou uvedeny intervaly spolehlivosti, hodnota t-statistiky pro test hypotézy, že daný parametr modelu AR(9) je nula a příslušná hladina pravděpodobnosti.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	0.1957	(0.0638,0.3276)	2.908	0.0036
2	0.1432	(0.0091,0.2773)	2.0934	0.0363
3	-0.0395	(-0.1752,0.0962)	-0.5699	0.5687
4	0.1044	(-0.0313,0.2401)	1.5078	0.1316
5	-0.0282	(-0.1648,0.1085)	-0.4039	0.6863
6	0.1128	(-0.0234,0.2490)	1.6229	0.1046
7	0.0590	(-0.0779,0.1959)	0.8445	0.3984
8	-0.0807	(-0.2170,0.0555)	-1.1611	0.2456
9	0.2216	(0.0871,0.3562)	3.2294	0.0012

S odhadem parametrů úzce souvisí přístup k detekci odlehlých pozorování založený na diagonálních prvcích projekční matice popsaný v teoretické části. Použitím

tohoto postupu dostaneme odlehlé vektory $\mathbf{z}_t = (X_{t-1}, \ldots, X_{t-9})$ pro indexy $t = 22, 24, 30, 31, 32, 33, 36, 37, 38, 71, 72, 73, 101, 102, 103, 104, 105, 106, 107$. Jako odlehlá pozorování můžeme identifikovat X_{29}, X_{70} a X_{100}. Tato pozorování odpovídají rokům 1799, 1840 a 1870.

U modelu AR(9) na pětiprocentní hladině pravděpodobnosti nezamítáme nulovost 6 z celkového počtu 9 parametrů (ani na desetiprocentní hladině to není o nic lepší). Bylo by vhodné pokusit se najít nějaký model, kdy T_t závisí pouze na nějaké podmnožině z posunutí $T_{t-1}, \ldots, T_{t-K}$, kde T_{t-K} je časově nejvzdálenější pozorování na kterém může T_t záviset. K výběru nejlepšího „podmnožinového" autoregresního modelu mezi velkým množstvím (2^K) modelů můžeme využít informační kritéria popsaná v teoretické části. Nejmenší AIC (-0.3414) a nejmenší HQ (-0.3484) pro $K = 15$ (mezi 32768 možnostmi) dostaneme pro následující model:

$$T_t = 0.1815T_{t-1} + 0.1339T_{t-2} + 0.1867T_{t-9} + 0.0963T_{t-13} + 0.1638T_{t-15} + \varepsilon_t, \tag{7.2}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 0.6795$. Portmanteau statistika $Q_{16} = 9.3707$, příslušná kritická hodnota na pětiprocentní hladině pravděpodobnosti je 33.1. Hodnota Box-Pierceovy statistiky Q_{16}^* je 9.8635, kritická hodnota je stejná jako pro portmanteau statistiku. Portmanteau ani Box-Pierceova statistika platnost ověřovaného modelu nezamítají. McLeod-Liova statistika $Q_{16}^{**} = 18.2443$ a příslušná kritická hodnota je 41.3, takže ani tento test nezamítá platnost ověřovaného modelu. Intervaly spolehlivosti pro jednotlivé parametry jsou uvedeny v následující tabulce:

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	0.1815	(0.0551,0.3078)	2.8148	0.0049
2	0.1339	(0.0053,0.2625)	2.0410	0.0413
9	0.1867	(0.0578,0.3155)	2.8398	0.0045
13	0.0963	(-0.0394,0.2274)	1.4387	0.1502
15	0.1638	(0.0312,0.2963)	2.4221	0.0154

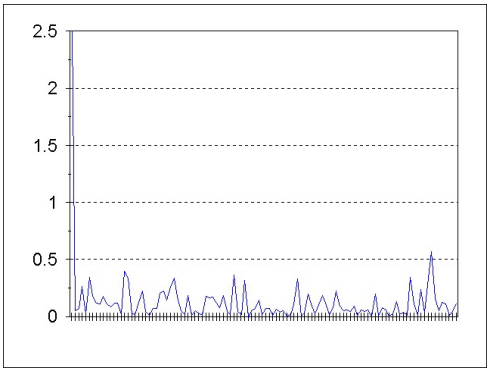
Pokud tento model porovnáme s modelem AR(9), zjistíme, že se snížil jak počet parametrů, tak i odhad rozptylu bílého šumu.

Stejným postupem pomocí kritéria BIC bychom dostali model

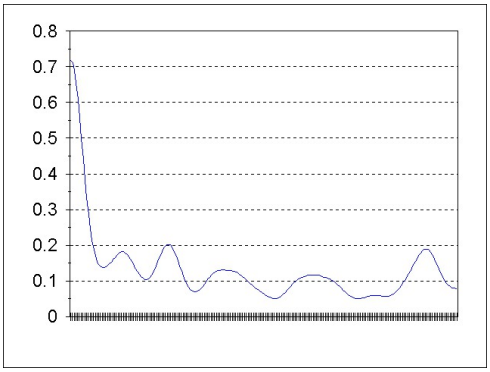
$$T_t = 0.2098T_{t-1} + 0.2157T_{t-9} + 0.1998T_{t-15} + \varepsilon_t, \tag{7.3}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 0.7006$. Portmanteau statistika ($Q_{16} = 14.5386$), Box–Pierceova statistika ($Q_{16}^* = 15.1571$), ani McLeod–Liova statistika ($Q_{16}^{**} = 19.0359$) platnost modelu nezamítají. U tohoto modelu se oproti modelu (7.2) o něco zvýšil odhad rozptylu bílého šumu, ale zato můžeme na pětiprocentní hladině zamítnout nulovost všech parametrů, viz následující tabulka.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	0.2098	(0.0853,0.3342)	3.3035	0.0010
9	0.2157	(0.0874,0.3439)	3.2962	0.0010
15	0.1998	(0.0692,0.3304)	2.9987	0.0027



Obrázek 7.4: Periodogram pro teploty



Obrázek 7.5: Odhad spektrální hustoty pro teploty

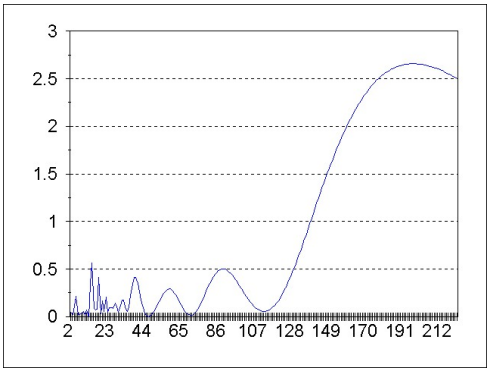
7.2 Spektrum

Spektrální analýza je vhodná zvláště pro zjisťování případné periodicity. Na obrázku 7.4 je periodogram pro teploty.

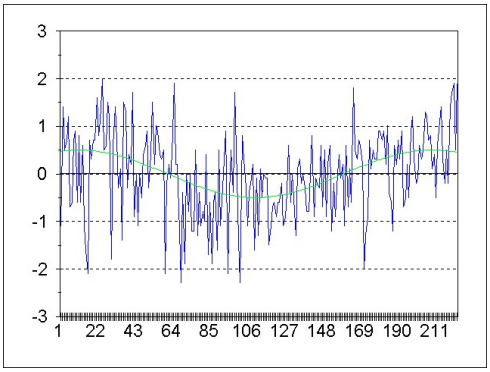
Největší normovaná hodnota periodogramu, která se používá jako testová statistika ve Fisherově testu je $W = 0.1717$ a odpovídá délce periody 222. Aproximace kritické hodnoty Fisherova testu je 0.0697 a Fisherův test tedy na pětiprocentní hladině pravděpodobnosti náhodnost dat zamítá. Pak jsem použil postup, který navrhl Whittle pro hledání dalších významných periodických složek, ale žádná další periodická složka už kritickou hodnotu Fisherova testu nepřekročila.

Hodnota statistiky T ze Siegelova testu je $T = 0.1298$, příslušná kritická hodnota je 0.0350 a Siegelův test nulovou hypotézu také zamítá.

Odhad spektrální hustoty získaný pomocí Parzenových vah je na obrázku 7.5. Vzhledem k tomu, že Fisherův test objevil významnou periodicitu a vzhledem k jeho konstrukci, kdy se vlastně testovala přítomnost periodických složek odpovídajících periodám $222, 111, \ldots$ (o chování periodogramu pro periody např. mezi 111 a 222 nemáme žádnou informaci),



Obrázek 7.6: Periodogram teplot pro délky period



Obrázek 7.7: Průměrné teploty a odhadnutá perioda

je vhodné spočítat periodogram odpovídající periodám $2, 3, \ldots, 222$ abychom mohli určit periodu, pro kterou skutečně periodogram nabývá svého maxima. Periodogram pro periody $2, \ldots, 222$ je na obrázku 7.6 a největší hodnoty nabývá pro periodu délky 197.

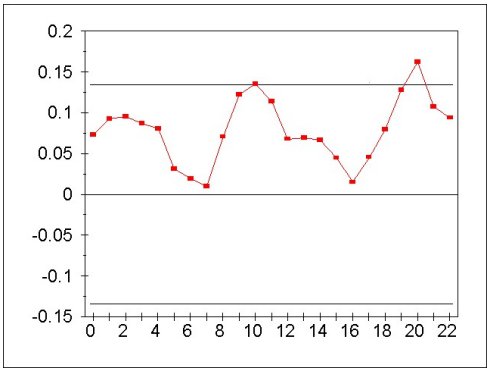
Metodou nejmenších čtverců jsem proložil daty sinusoidu s periodou 197

$$T_t = 0.1707 \sin(2\pi t/197) + 0.4682 \cos(2\pi t/197) + z_t, \tag{7.4}$$

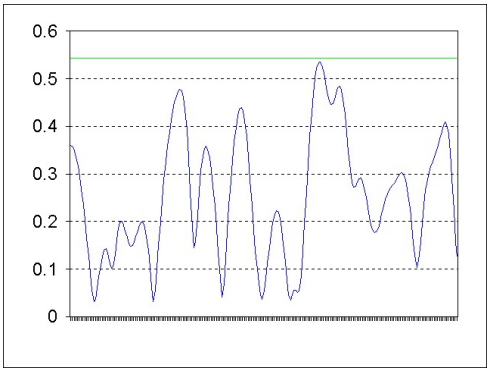
kde z_t jsou odchylky T_t od proložené sinusoidy, které se pokusím dále modelovat. Odhadnutá sinusoida je i s průměrnými ročními teplotami na obrázku 7.7. Intervaly spolehlivosti pro odhadované parametry z programu SOLO jsou uvedeny v následující tabulce:

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
sin	0.1707	(0.0095,0.3318)	2.0761	0.0379
cos	0.4682	(0.3177,0.6187)	6.0985	0.0000

Pro časovou řadu z_t jsem se pokusil najít vhodný autoregresní model. Podle kritéria BIC použitého na všechny „podmnožinové“ modely pro $K = 15$ je nejrozumnější řadu z_t dále autoregresně nemodelovat a prohlásit ji za bílý šum s rozptylem 0.6889. Pro tento



Obrázek 7.8: Vzájemná korelační funkce pro teploty a Wolfova čísla



Obrázek 7.9: Koherenční diagram pro teploty a Wolfova čísla

model je portmanteau statistika $Q_{16} = 20.4186$, Box–Pierceova statistika $Q^*_{16} = 21.4676$ a McLeod–Liova statistika $Q^{**}_{16} = 22.8730$. Ani jedna statistika nepřekračuje příslušnou kritickou hodnotu.

Kritéria AIC a HQ se pro časovou řadu z_t shodla na modelu

$$z_t = 0.1085 z_{t-1} - 0.1071 z_{t-8} + 0.1324 z_{t-9} - 0.1334 z_{t-12} + 0.1133 z_{t-15} + \varepsilon_t, \qquad (7.5)$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma^2_\varepsilon = 0.6415$. Pro tento model je portmanteau statistika $Q_{16} = 5.1423$, Box–Pierceova statistika $Q^*_{16} = 5.3201$ a McLeod–Liova statistika $Q^{**}_{16} = 17.5816$. Opět ani jedna statistika nepřekračuje příslušnou kritickou hodnotu rozdělení χ^2. Intervaly spolehlivosti pro autoregresní parametry jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
1	0.1085	(-0.0215,0.2385)	1.6361	0.1018
8	-0.1071	(-0.2393,0.0251)	-1.5877	0.1124
9	0.1324	(-0.0007,0.2655)	1.9486	0.0513
12	-0.1334	(-0.2663,-0.0004)	-1.9672	0.0492
15	0.1133	(-0.0198,0.2464)	1.6690	0.0951

7.3 Závislost na Wolfových číslech

Odhad vzájemné korelační funkce $r_{TW}(k)$ pro teploty a Wolfova čísla je na obrázku 7.8. Vzhledem k tomu, že závislost sluneční aktivity na průměrných teplotách v Klementinu by se velice špatně zdůvodňovala, počítal jsem hodnoty $r_{TW}(k)$ jenom pro $k \geq 0$. Protože máme k dispozici Wolfova čísla už od roku 1749, můžeme $r_{TW}(k)$ pro $k = 0, \ldots, 22$ počítat jako

$$r_{TW}(k) = \frac{\sum_{i=1}^{n} T_i W_{i-k}}{\sqrt{\sum_{i=1}^{n} T_i^2 \sum_{i=1}^{n} W_{i-k}}}.$$

Jednoduchý odhad směrodatné odchylky $r_{TW}(k)$ je $\hat{\sigma}_k = 1/\sqrt{n}$, kde n je počet pozorování, v našem případě 222. Porovnáním $|r_{TW}(k)|$ s dvojnásobkem této směrodatné odchylky získáme jednoduchý test významnosti dané hodnoty vzájemné korelační funkce. Meze získané touto metodou jsou na obrázku 7.8 také vyznačeny a je vidět, že jsou překročeny až pro příliš velké zpoždění, než aby se dala tato závislost rozumně interpretovat.

V rámci spektrální analýzy můžeme hodnotit závislost časových řad pomocí koherenčního koeficientu (obr. 7.9), který nikde nepřekračuje kritickou hodnotu založenou na Goodmanově aproximaci.

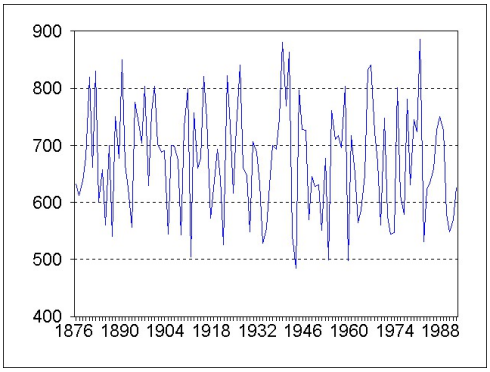
Kapitola 8

Srážky

V následující tabulce jsou uvedeny roční úhrny srážek v Klementinu naměřené v letech 1876 až 1992. Stejná data jsou graficky znázorněna na obrázku 8.1.

1876	632	613	637	679	819	662	830	602	657	560
1886	700	541	750	677	850	665	625	557	775	738
1896	704	803	630	758	804	702	688	691	545	700
1906	699	674	543	732	799	505	757	660	677	821
1916	746	573	624	693	641	526	822	734	616	730
1926	840	662	647	549	706	688	632	529	555	631
1936	700	693	748	880	769	863	542	485	797	728
1946	726	570	646	628	631	551	678	499	761	710
1956	717	696	804	498	717	650	565	587	648	833
1966	841	724	670	560	747	576	545	547	801	612
1976	579	781	631	745	724	886	531	623	635	658
1986	727	751	726	579	549	570	627			

Nejdříve jsem na data použil všechny testy náhodnosti popsané v teoretické části.



Obrázek 8.1: Roční úhrny srážek v Klementinu

Celkem je v časové řadě ročních úhrnů srážek 117 pozorování. Vedle sebe nikde nejsou stejné hodnoty. Střední hodnota počtu bodů růstu je 58, rozptyl je 9.83. Skutečný počet bodů růstu je 56. Testová statistika $\frac{|56-58|}{\sqrt{9.83}} = |-0.64| < 1.96$ a na pětiprocentní hladině pravděpodobnosti hypotézu náhodnosti dat nezamítáme.

Počet bodů zvratu je 73. Střední počet bodů zvratu je 76 , rozptyl je 20.48. Testová statistika $\frac{|73-76|}{\sqrt{20.48}} = |-0.81|$ je menší než kritická hodnota normálního rozdělení na pětiprocentní hladině pravděpodobnosti 1.96 a náhodnost dat tedy nezamítá ani test založený na bodech zvratu.

Spearmanův koeficient pořadové korelace je -0.0848 a opět není překročena kritická hodnota na pětiprocentní hladině pravděpodobnosti.

Při mediánovém testu leží nad mediánem 57 pozorování, pod mediánem 58 pozorování a na mediánu 2 pozorování (medián je 677). Počet vytvořených sérií je 51. Testová statistika nabývá hodnoty -1.22 a není překročena kritická hodnota.

Test pomocí odhadu rozptylu ze sousedních pozorování náhodnost dat na pětiprocentní hladině pravděpodobnosti také nezamítá (hodnota statistiky d je 1.94).

Výsledky Fisherova a Siegelova testu jsou podrobně uvedeny dále, ale ani tyto dva testy náhodnost dat na pětiprocentní hladině pravděpodobnosti nezamítají.

Žádný test náhodnosti popsaný v teoretické části na pětiprocentní hladině pravděpodobnosti nezamítl náhodnost dat. Zatím se časová řada ročních úhrnů srážek jeví nejspíše jako posloupnost nezávislých náhodných veličin.

Dále budu pracovat pouze s vycentrovanými daty (od dat odečtu jejich výběrový průměr, abych mohl předpokládat, že analyzovaná časová řada má nulovou střední hodnotu). Výběrový průměr této časové řady je 674.1025. Centrovanou časovou řadu ročních úhrnů srážek v Klementinu od roku 1876 do roku 1992 budu v dalším textu značit jako $S_1, \ldots, S_{117}$.

8.1 Autoregrese

I když výsledky testů náhodnosti nejsou příliš povzbudivé, pokusil jsem se nalézt autoregresní model. Odhady autokorelační a parciální autokorelační funkce jsou uvedeny v následující tabulce:

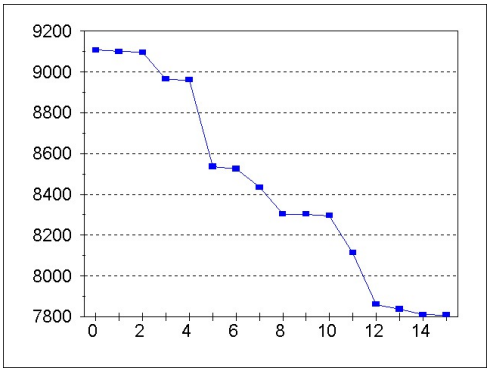
	1	2	3	4	5	6	7	8
autokorelace	0.028	-0.016	-0.120	-0.024	-0.206	-0.022	-0.078	-0.061
parciální autokorelace	0.028	-0.016	-0.119	-0.018	-0.212*	-0.029	-0.099	-0.119

	9	10	11	12	13	14	15	16
autokorelace	0.026	0.110	-0.087	-0.109	0.088	0.001	0.005	0.126
parciální autokorelace	0.004	0.034	-0.137	-0.154	0.060	-0.053	-0.021	0.102

Významnost hodnot parciální autokorelační funkce můžeme opět posoudit pomocí Quenouilleovy aproximace pro směrodatnou odchylku parciální autokorelační funkce. Významnost hodnot r_{kk} určujeme porovnáním s číslem $2/\sqrt{117} = 0.185$. Významná hodnota $r_{5,5}$ je označena hvězdičkou.

Odhady rozptylu bílého šumu po autoregresi řádu k jsou na obrázku 8.2. Odhad rozptylu bílého šumu se zmenší pro řády autoregrese 5 a 12.

Jako další nástroj pro určení řádu autoregrese jsem použil informační kritéria AIC, BIC a HQ, popsaná v teoretické části. Všechna kritéria se shodla na řádu autoregrese 0 a zřejmě



Obrázek 8.2: Odhad rozptylu bílého šumu pro srážky jako AR(k)

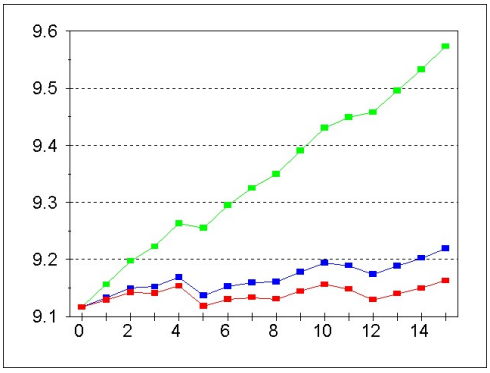
tedy nemá smysl pokoušet se hledat vhodný autoregresní model. Informační kritéria AIC, BIC a HQ jsou uvedena v následující tabulce, nejmenší hodnoty jednotlivých kritérií jsou označeny hvězdičkou. Tytéž hodnoty jsou graficky znázorněny na obrázku 8.3.

řád AR modelu	odhad σ_ε^2	AIC	BIC	HQ
0	9106.6561	9.1168*	9.1168*	9.1168*
1	9099.4096	9.1331	9.1567	9.1293
2	9096.9006	9.1499	9.1971	9.1424
3	8964.5496	9.1523	9.2231	9.1410
4	8960.3103	9.1689	9.2634	9.1539
5	8534.9392	9.1374	9.2554	9.1186
6	8524.2425	9.1532	9.2949	9.1307
7	8433.9602	9.1597	9.3249	9.1334
8	8302.6842	9.1611	9.3499	9.1310
9	8302.6665	9.1782	9.3906	9.1444
10	8294.2883	9.1943	9.4303	9.1567
11	8113.0475	9.1893	9.4490	9.1480
12	7859.4656	9.1746	9.4579	9.1295
13	7836.9485	9.1888	9.4957	9.1400
14	7810.1466	9.2025	9.5330	9.1499
15	7808.0723	9.2193	9.5734	9.1630

Opět jsem se pokusil najít autoregresní model, ve kterém S_t závisí pouze na nějaké podmnožině posunutých pozorování $S_{t-1}, \ldots, S_{t-15}$. Kritérium AIC jako nejvhodnější (mezi $2^{15} = 32768$ možnostmi) vybralo následující model:

$$S_t = -0.2318 S_{t-5} - 0.1524 S_{t-12} + \varepsilon_t, \tag{8.1}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 8519.7$. Pro tento model je portmanteau statistika $Q_{12} = 6.9819$, Box–Pierceova statistika $Q_{12}^* = 7.5064$ a McLeod–Liova statistika



Obrázek 8.3: AIC, BIC a HQ pro stanovení řádu autoregrese srážek

$Q_{12}^{**} = 9.7027$. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 31.4, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Ani jeden test nezamítá platnost zkonstruovaného modelu.

Intervaly spolehlivosti pro autoregresní parametry získané z programu SOLO jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
5	-0.2318	(-0.4153,0.0483)	-2.5026	0.0137
12	-0.1524	(-0.3433,0.0386)	-1.5809	0.1166

Kritérium BIC je nejmenší pro model

$$S_t = -0.2141 S_{t-5} + \varepsilon_t, \tag{8.2}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 8704.9$. Hodnota portmanteau statistiky Q_{12} pro tento model je $Q_{12} = 9.6247$, Box–Pierceova statistika Q_{12}^* nabývá hodnoty $Q_{12}^* = 10.4765$ a McLeod–Liova statistika Q_{12}^{**} se rovná 6.0807. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 33.1, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Na pěti-procentní hladině pravděpodobnosti tedy platnost tohoto modelu nezamítáme.

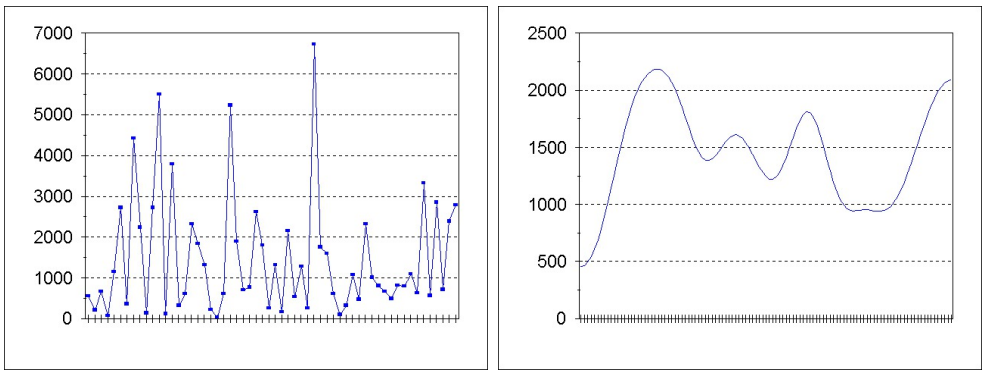
Intervaly spolehlivosti pro autoregresní parametry získané z programu SOLO jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
5	-0.2141	(-0.3975,-0.0308)	-2.3139	0.0224

Kritérium HQ je nejmenší pro model

$$\begin{aligned} S_t = -0.1638 S_{t-3} - 0.2593 S_{t-5} - 0.1250 S_{t-7} - 0.1389 S_{t-8} - \\ -0.1265 S_{t-11} - 0.1743 S_{t-12} + \varepsilon_t, \end{aligned} \tag{8.3}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 8001.4$. Portmanteau statistika Q_{12} nabývá hodnoty $Q_{12} = 2.6208$, Box–Pierceova statistika $Q_{12}^* = 2.8063$ a McLeod–Liova statistika



Obrázek 8.4: Periodogram a odhad spektrální hustoty pro srážky

$Q_{12}^{**} = 8.0102$. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 24.1, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Ani platnost tohoto modelu tedy na pětiprocentní hladině pravděpodobnosti nezamítáme.

Pouze u jednoho parametru nezamítáme na pětiprocentní hladině pravděpodobnosti jeho nulovost. Proto se mi tento model nezdá příliš vhodný. Nejmenší hladina pravděpodobnosti, na které ještě zamítáme nulovost parametrů a intervaly spolehlivosti jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
3	-0.1638	(-0.3481,0.0204)	-1.7619	0.0808
5	-0.2593	(-0.4420,-0.0766)	-2.8120	0.0058
7	-0.1250	(-0.3102,0.0601)	-1.3381	0.1836
8	-0.1389	(-0.3264,0.0486)	-1.4681	0.1449
11	-0.1265	(-0.3118,0.0587)	-1.3538	0.1786
12	-0.1743	(-0.3677,0.0191)	-1.7855	0.0769

8.2 Spektrum

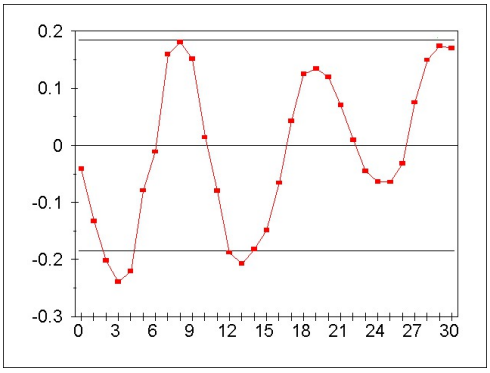
Periodogram pro roční úhrny srážek je na obrázku 8.4. Nejvyšší hodnota normovaného periodogramu je dosažena pro periodu 3.25, $W = 0.0793$. Kritická hodnota Fisherova testu na pětiprocentní hladině pravděpodobnosti (0.1212) není překročena a hypotézu náhodnosti dat nezamítáme. Hodnota Siegelovy statistiky T je 0.0065 a kritická hodnota na pětiprocentní hladině pravděpodobnosti (0.0554) není překročena ani v tomto případě.

Odhad spektrální hustoty pomocí Parzenových vah je na obrázku 8.4.

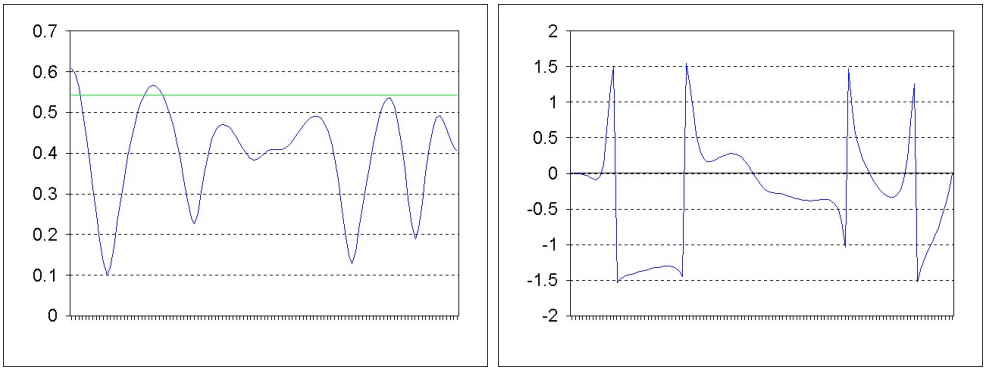
8.3 Závislost na Wolfových číslech a teplotách

Závislost na Wolfových číslech

Vzájemná korelační funkce $r_{SW}(k)$ pro srážky a Wolfova čísla je na obrázku 8.5 a má na první pohled velice pravidelný průběh. Zajímavé jsou především nízké hodnoty vzájemné korelační



Obrázek 8.5: Vzájemná korelační funkce pro srážky a Wolfova čísla



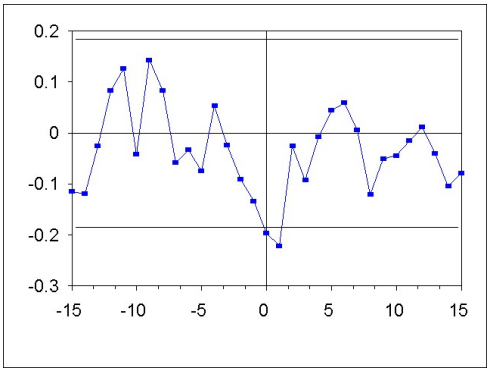
Obrázek 8.6: Koherenční a fázový diagram pro Wolfova čísla a srážky

funkce pro $k = 2, 3, 4$, které svědčí o poněkud opožděné závislosti srážek na Wolfových číslech.

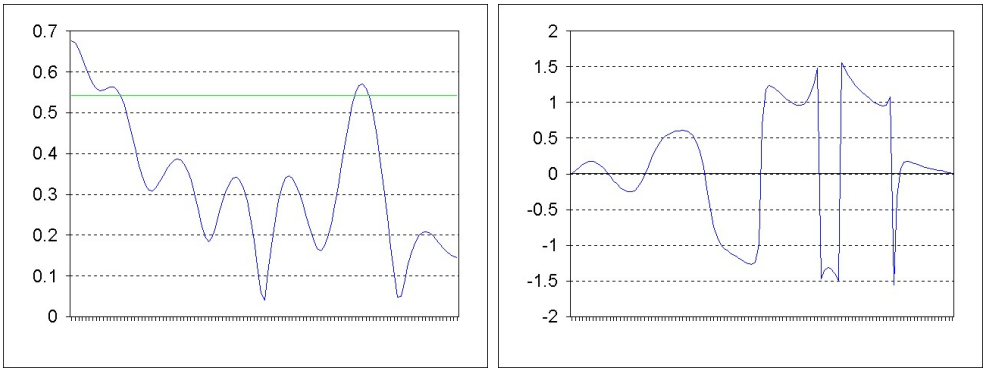
Příslušný koherenční diagram je na obrázku 8.6. Kritická hodnota založená na Goodmanově aproximaci je překročena pro délky period větší než 110 a pro délky period zhruba mezi 10.5 a 8.8. Největší hodnota koherenčního diagramu pro délky period mezi 10.5 a 8.8 je dosažena pro délku periody 9.57. Odpovídající hodnota fáze $\Phi_{WS}(\lambda)$je -1.3381, což odpovídá časovému zpoždění srážek za Wolfovými čísly $(-1.3381 + \pi) \cdot 9.57/2\pi = 2.7469$. Významná hodnota koherenčního koeficient pro délku periody i odpovídající časové zpoždění v rámci této periody pěkně odpovídají průběhu vzájemné korelační funkce $r_{SW}(k)$. Fázový diagram pro Wolfova čísla a srážky je na obrázku 8.6.

Závislost mezi srážkami a teplotami

Vzájemná korelační funkce mezi srážkami a teplotami $r_{ST}(k)$ je na obrázku 8.7. Významně se liší od nuly pro $k = 0$ a $k = 1$. Tento výsledek se dá interpretovat tak, že hodně srážek snižuje v daném roce průměrnou teplotu a vyšší průměrné teploty snižují v příštím roce množství srážek. Pro $k < 0$ jsou hodnoty vzájemné korelační funkce malé, takže průměrné roční teploty na množství srážek v minulých letech zřejmě nezávisí.



Obrázek 8.7: Vzájemná korelační funkce pro srážky a teploty



Obrázek 8.8: Koherenční a fázový diagram pro teploty a srážky

Koherenční koeficient (obr. 8.8) $C_{TS}(\lambda)$ vychází významně pro délky period větší než 16 a pro délky period mezi 2.62 a 2.72. Největší hodnota koherenčního diagramu pro délky period mezi 2.62 a 2.72 je dosažena pro délku periody 2.65. Odpovídající hodnota fáze $\Phi_{TS}(\lambda)$je 1.1878, což odpovídá časovému zpoždění srážek za teplotami $1.1878 \cdot 2.65/2\pi = 0.5010$. Toto časové zpoždění pěkně odpovídá vzájemné korelační funkci $r_{ST}(k)$, která také ukazuje obdobné zpoždění. Fázový diagram pro teploty a srážky je na obrázku 8.8.

Odhad modelu

Model, který bere do úvahy závislost na teplotách a Wolfových číslech jsem pomocí kritéria AIC odhadl jako

$$\begin{aligned} S_t = &-0.2033 S_{t-3} - 0.2142 S_{t-5} - 0.1636 S_{t-8} - 0.1779 S_{t-11} \\ &-0.1642 S_{t-12} - 24.4743 T_{t-1} - 0.5198 W_{t-4} + \varepsilon_t, \end{aligned} \tag{8.4}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 7230.45$. Pro tento model je portmanteau statistika $Q_{12} = 8.46412$, Box–Pierceova statistika $Q_{12}^* = 9.0728$ a McLeod–Liova statistika

$Q_{12}^{**} = 9.5907$. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 22.1, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Ani jedna statistika nepřekročila kritickou hodnotu na pětiprocentní hladině pravděpodobnosti a platnost tohoto modelu tedy nezamítáme. Intervaly spolehlivosti pro parametry tohoto modelu jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
S_{t-3}	-0.2033	(-0.3806,-0.0261)	-2.2732	0.0250
S_{t-5}	-0.2142	(-0.3899,-0.0384)	-2.4150	0.0174
S_{t-8}	-0.1636	(-0.3432,0.0160)	-1.8054	0.0737
S_{t-11}	-0.1779	(-0.3579,0.0022)	-1.9577	0.0528
S_{t-12}	-0.1642	(-0.3478,0.0194)	-1.7724	0.0791
T_{t-1}	-24.4743	(-45.8005,-3.1481)	-2.2743	0.0249
W_{t-4}	-0.5198	(-0.9050,-0.1347)	-2.6749	0.0086

Pomocí kritéria BIC jsem dostal model

$$S_t = -0.2045 S_{t-5} - 26.2091 T_{t-1} + \varepsilon_t, \tag{8.5}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 8308.58$. Hodnota portmanteau statistiky pro tento model je $Q_{12} = 14.0720$, Box–Pierceova statistika Q_{12}^{*} je 15.2433 a McLeod–Liova statistika $Q_{12}^{**} = 7.6440$. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 31.4, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Platnost tohoto modelu na pětiprocentní hladině pravděpodobnosti nezamítáme. V následující tabulce jsou uvedeny intervaly spolehlivosti pro parametry.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
S_{t-5}	-0.2045	(-0.3845,-0.0245)	-2.2502	0.0263
T_{t-1}	-26.2091	(-48.2302,-4.1879)	-2.3575	0.0201

Kritérium HQ jako nejlepší určilo skoro stejný model jako kritérium AIC:

$$\begin{aligned} S_t = &-0.1966 S_{t-3} - 0.2078 S_{t-5} - 0.1616 S_{t-8} - 0.2032 S_{t-11} \\ &-0.1611 S_{t-12} - 29.4081 T_{t-1} + 14.9518 T_{t-2} - 0.5722 W_{t-4} + \varepsilon_t, \end{aligned} \tag{8.6}$$

kde $\{\varepsilon_t\}$ je bílý šum s rozptylem $\sigma_\varepsilon^2 = 7125.69$. Pro tento model je portmanteau statistika $Q_{12} = 7.2270$, Box–Pierceova statistika $Q_{12}^{*} = 7.7260$ a McLeod–Liova statistika $Q_{12}^{**} = 10.4361$. Kritická hodnota pro portmanteau a Box–Pierceovu statistiku je 20.0, kritická hodnota pro McLeod–Liovu statistiku je 34.8. Na pětiprocentní hladině pravděpodobnosti tedy platnost ověřovaného modelu nezamítáme. Konfidenční intervaly pro parametry tohoto modelu jsou uvedeny v následující tabulce.

parametr	odhad	95% interval spolehlivosti	t-statistika	hladina pravděpodobnosti
S_{t-3}	-0.1966	(-0.3737,-0.0196)	-2.2008	0.0299
S_{t-5}	-0.2078	(-0.3833,-0.0322)	-2.3454	0.0208
S_{t-8}	-0.1616	(-0.3408,0.0175)	-1.7879	0.0766
S_{t-11}	-0.2032	(-0.3871,-0.0193)	-2.1898	0.0307
S_{t-12}	-0.1611	(-0.3442,0.0221)	-1.7429	0.0842
T_{t-1}	-29.4081	(-52.0378,-6.7784)	-2.5756	0.0113
T_{t-2}	14.9518	(-8.4582,38.3618)	1.2659	0.2083
W_{t-4}	-0.5722	(-0.9650,-0.1794)	-2.8873	0.0047

Obsah

1 Základní pojmy **2**
1.1 Příklady stacionárních posloupností 4

2 Testy náhodnosti **5**

3 Autoregresní posloupnosti **10**
3.1 Odhad parametrů a odlehlá pozorování 11
3.2 Určení řádu autoregresního modelu 13
3.3 Ověřování modelu 15

4 Spektrální analýza **17**
4.1 Teorie 17
4.2 Periodogram a testování periodicity 19
4.3 Odhad spektrální hustoty 21

5 Testování nezávislosti mezi časovými řadami **22**
5.1 Metody založené na korelaci 22
5.2 Spektrální metody 25

6 Wolfova čísla **29**

7 Teploty **36**
7.1 Autoregrese 37
7.2 Spektrum 41
7.3 Závislost na Wolfových číslech 44

8 Srážky **45**
8.1 Autoregrese 46
8.2 Spektrum 49
8.3 Závislost na Wolfových číslech a teplotách 49

Literatura

[1] Anděl, J.: Statistická analýza časových řad. Praha, SNTL 1976

[2] Anděl, J.: Matematická statistika. Praha, SNTL/Alfa 1978

[3] Anděl, J.: Některé postupy užívané při hodnocení časových řad. Praha, MFF UK 1981

[4] Box, G.E.P.– Jenkins, G.M.: Time Series Analysis, Forecasting and Control. San Francisco, Holden Day 1970

[5] Cipra, T.: Analýza časových řad s aplikacemi v ekonomii. Praha, SNTL/Alfa 1986

[6] Cramér, H.: Mathematical Methods of Statistics. Princeton, Princeton Univ. Press 1946

[7] Fabian, V.: Základní statistické metody. Praha, ČSAV 1963

[8] Goodman, N.R.: On the joint estimation of the spectra, cospectrum and quadrature spectrum of a two dimensional stationary Gaussian process. Scientific Paper No. 10, New York University, Engineering Statistics Laboratory 1957

[9] Goodman, N.R.: Statistical analysis based on a certain multivariate complex Gaussian distribution. *Annals of Mathematical Statistics*, **34**, 152–177, 1963

[10] Hannan, E.J.– Quinn, B.G.: The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, **41**, 190–195, 1979

[11] Haugh, L.D.: Checking the independence of two covariance stationary time series: a univariate residual cross-correlation approach.*Journal of the American Statistical Association*, **71**, 378–385, 1976

[12] Janko, J.: Statistické tabulky. NČSAV, Praha 1958

[13] Kendall, M.: Time-series. London, Griffin 1948

[14] Likeš, J. – Laga, J.: Základní statistické tabulky. Praha, SNTL 1978

[15] Ljung, G.M.– Box G.E.P.: On a measure of lack in time series models. *Biometrika*, **65**, 293–303, 1978

[16] McLeod, A.I.– Li, W.K.: Diagnostic checking ARMA time series models using squared-residual autocorrelations. *J. Time Ser. Anal.*, **4**, 269–273, 1983

[17] Moran, P.A.P.: Some experiments on the prediction of sunspot numbers. *Journal of the Royal Statistical Society B*, **16**, 112–117, 1954

[18] Morris, M.J.: Forecasting the Sunspot Cycle. *Journal of the Royal Statistical Society A*, **140**, Part 4, 437–468, 1977

[19] Quenouille, M.H.: Aproximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, **11**, 68–84, 1949

[20] Rissanen, J.: Modelling by shortest data description.*Automatica*, **14**, 465–471, 1978

[21] Schaerf, M.C.: Estimation of the covariance and autoregressive structure of a stationary time series. Technical Report, Department of Statistics, Stanford University, Stanford, California, 1964

[22] Schwarz, G.: Estimating the dimension of a model.*Annals of Statistics*, **6**, 461–464, 1978

[23] Siegel, A.F.: Testing for Periodicity in a Time Series. *Journal of the American Statistical Association*, **75**, 345–348, 1980

[24] Tong, H.: Some Comments on the Canadian Lynx Data. *Journal of the Royal Statistical Society A*, **140**, Part 4, 432–436, 1977

[25] Tong, H.: Non-linear Time Series. Oxford, Clarendon Press 1990

[26] Waldmeier, M.: Sonnenfleckenkurven und die Methode der Sonnenaktivitäts Prognose. Astronomische Mitteilungen der Eidgenossischen Sternwarte Zurich, No. 286, 1968

[27] Whittle, P.: The simultaneous estimation of time series, harmonic components and covariance structure. *Trabajos de Estadistica*, **3**, 43–57, 1952

[28] Wolfowitz, J.: Asymptotic distribution of runs up and down. *Annals of Mathematical Statistics*, **15**, 163–172, 1944

[29] Yule, G.U.: On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot series. *Phil. Trans. Roy. Soc., Series A*, **226**, 267–298, 1927